تمثيل المعرفة

واسترجاع المعلومات الرقمية

أ.د. خالد عبد الفتاح محمد

أستاذ علم المعلومات وإدارة المعرفة

مدير مكتبة دبي الرقمية وحلول المعرفة



تمثيل المعرفة

واسترجاع المعلومات الرقمية

تمثيل المعرفة

واسترجاع المعلومات الرقمية

أ.د. خالد عبد الفتاح محمد

أستاذ علم المعلومات وإدارة المعرفة

مدير مكتبة دبي الرقمية وحلول المعرفة



Qindeel | قنديل

Knowledge Representation
and Digital Information Retrieval

Prof. Khaled Abd Elfatah Mohamed

تمثيل المعرفة
واستترجاع المعلومات الرقمية

أ. د خالد عبد الفتاح محمد

موافقة «**المجلس الوطني للإعلام**» في دولة الإمارات العربية المتحدة
رقم: MC-10- 01-3557795 تاريخ 2019/1/20

ISBN: 978 - 9948 - 38 - 751 - 0



للطباعة والنشر والتوزيع
Printing, publishing & Distribution

ص. ب: 47417 شارع الشيخ زايد
دبي – دولة الإمارات العربية المتحدة
البريد الإلكتروني: info@qindeel.ae
الموقع الإلكتروني: www.qindeel.ae

الطبعة الأولى: نيسان / إبريل 2019 م – 1440 هـ

المحتويات

الفصل الرابع

الفصل الخامس

الفصل السادس

الفصل السابع

آليات الاسترجاع وتمثيل الاستفسارات 215

الفصل التاسع

نماذج استرجاع المعلومات **297**

الفصل العاشر

الإهداء

إلى من غابت عنهم الأعين وسكنوا القلب والعقل،

أبي الغالي وأستاذي أ.د حشمت قاسم

رحمهما الله وطيب ثراهما

وإلى أمي الغالية التي بها نغنى ونستغني.

مقدمة

إن لم تكن صاحب فضل

فلا تنس للناس أفضالهم

فجُدْ بنسب الجميل لأهله

واذكر لكل كريم خصاله

لقد كان أستاذي الكبير العالم الجليل، أ.د حشمت قاسم، رائد علم المعلومات وأفضل من كتب وترجم مؤلفات عالمية في مجال استرجاع المعلومات، الدافع الأكبر نحو تأليف هذا الكتاب. فقد راجع أول بحث أعددته باللغة الإنجليزية وآخر بالعربية، وكانا في مجال استرجاع المعلومات، فحفزني إلى ضرورة ترجمة أو تأليف كتاب في مجال استرجاع المعلومات.

عكفت أكثر من خمس سنوات على تأليف هذا الكتاب. طالعت خلالها وتابعت ما يحدث في هذا المجال من تطورات لم يتسع الكتاب لعرضها بالكامل. وأحسبه قد بدأ من النقطة التي توقف عندها آخر كتاب في هذا المجال ترجمه أستاذنا الفاضل أ.د حشمت قاسم والذي كان بعنوان «أساسيات استرجاع المعلومات». فوجدت أنه من الضروري أن يكون هناك كتاب يكمل ما حدث من تطورات في البيئة الرقمية التي شهدت ظهور آليات وأدوات جديدة لمعالجة واسترجاع المعلومات. وقد كانت أيضاً كلمات أستاذي رحمة الله عليه دافعاً ومحفزاً لإصدار الكتاب.

وقد تم بناء الهيكل العام لهذا الكتاب من منطلق التعامل مع قضايا تمثيل المعرفة ومعالجتها واسترجاعها على مستويين أساسيين هما: طرق المعالجة، والتوجهات الحديثة التى تناولتها الدراسات التى تم نشرها فى خلال العقدين الأول والثاني من القرن الجديد.

تم استخدام مصطلح تمثيل المعرفة في هذا الكتاب إشارة إلى المعنى العام للمعرفة الذي يتضمن البيانات والمعلومات والمعرفة. لذلك بدأ الكتاب بعرض لهرم المعرفة ومكوناته.

ويشتمل الكتاب على أحد عشر فصلاً، تناول كل فصل من هذه الفصول قضية أساسية من قضايا تمثيل المعرفة واسترجاعها. ويتم استخدام مصطلح المعرفة هنا على اتساعه بما يتضمنه من بيانات ومعلومات.

وقد بدأ الكتاب في **الفصل الأول** بعرض لقضايا تمثيل المعرفة من حيث المفاهيم والتعريفات الأساسية، وتطور آليات معالجة المعرفة وتمثيلها واسترجاعها.

تناول **الفصل الثاني** مشكلة تمثيل واسترجاع المعلومات بشقيها الرياضي الذي يركز على قياس كفاءة النظم وإمكانيات الاسترجاع، والإجرائي الذي يستعرض المكونات الأساسية لأي نظام لتمثيل المعرفة واسترجاع المعلومات وتحديات التمثيل والاسترجاع.

واستعرض **الفصل الثالث** طرق تمثيل المعرفة التي تتضمن خمس طرق أساسية هي: التكشيف، التصنيف أو التقسيم إلى فئات، التوسيم الاجتماعي، التلخيص، الملخص الوافي للموقع.

تناول **الفصل الرابع** مصادر البيانات بنظم تمثيل المعرفة والتي تأتي من ثلاثة مصادر أساسية هي البيانات والميتاداتا والنصوص الكاملة أو الكيانات الرقمية الكاملة. وقد عرض الفصل آليات هيكلة البيانات من خلال استخدام الميتاداتا وإجراءات معالجة الكيانات الرقمية وما تتضمنه من نصوص كاملة.

وركز **الفصلان الخامس والسادس** على مناقشة قضية اللغة ودورها في تمثيل واسترجاع المعرفة بمفهومها الواسع. وقد عرض الفصل الخامس أهم آليات تكويد المعرفة سواء من خلال آليات التصنيف الذي يستخدم دلالات رمزية أو من خلال لغات التكشيف الاصطناعية والطبيعية وأثر كل منهما في بنية النظم وإجراءات الاسترجاع. كما تم عرض تطور لغات التكشيف والتحديات التي تعالجها تلك اللغات كأدوات لتمثيل المعرفة. كما تم عرض لغات التكشيف في البيئة الرقمية بأنواعها المختلفة.

الفصل السابع تناول آليات البحث واسترجاع المعلومات والاعتبارات التي يجب مراعاتها عند إجراء عمليات البحث عن المعلومات، والتي تشمل تمثيل وصياغة الاستفسارات، إجراءات البحث وآلياته المختلفة سواء من حيث طريقة البحث أو حقول البحث. كما يعرض الفصل أساليب اختيار آلية البحث الملائمة إلى جانب معايير تقييم النتائج.

استعرض **الفصل الثامن** أساليب الاسترجاع التي تشمل ثلاثة أساليب أساسية هي: البحث، التصفح، والنموذج الهجين من البحث والتصفح. ويعالج هذا الفصل الأساليب الثلاثة المستخدمة في استرجاع المعلومات من حيث الملامح والتطبيقات والمزايا والعيوب.

وركز **الفصل التاسع** على عرض نماذج استرجاع المعلومات، التي تعتمد في الأساس على نظم المضاهاة والمطابقة بين المصطلحات، فاستعرض أساليب المضاهاة المختلفة، ثم النماذج الثلاثة الأساسية وهي النموذج البوليني، نموذج الفراغ الاتجاهي، النموذج الاحتمالي. واختتم الفصل بعرض لآليات الدمج بين النماذج لتوسيع إمكانيات نظم استرجاع المعلومات، والذي يتضمن النموذج البوليني الموسع ونموذج المجموعة الضبابية.

الفصلان العاشر والحادي عشر ركزا على الاسترجاع في بيئة الويب من خلال

استعراض ملامح بيئة الويب وتطور آليات الاسترجاع وأنواعها التي تضمنت الإبحار، التصفح، البحث مع التركيز على محركات البحث ومكوناتها وأنواع الزواحف وآليات عملها، ثم ما وراء المحركات وبوابات الويب وأنواعها. وركز الفصل الحادي عشر على عرض لمراجعة علمية تفصيلية للدراسات المتعلقة بتمثيل المعرفة بمحركات البحث وآليات تكشيفها وفرزها في بيئة الويب. وركز بصفة أساسية على المنهجيات والقياسات المتبعة في دراسات الويب. وقد تم تقسيم الدراسات إلى دراسات واقعية تعمل في البيئات التشغيلية ودراسات معملية تتم في المختبرات وفي بيئات اصطناعية، ثم تناول الفصل آليات التكشيف وطرق دراستها. وعرض لكل السبل الممكنة لدفع النتائج وترقيتها بمحركات البحث، إلى جانب عرض لطبيعة المشكلات التي تتناولها الدراسات بغرض توضيح اتجاهات الإنتاج الفكري في هذا المجال إلى جانب طبيعة المناهج والأساليب المتبعة في دراسة تلك المشكلات. وهذا الفصل على وجه الخصوص يعد أداة تمكن الباحثين من التعرف إلى طرق وأساليب إجراء دراسات الويب بصفة عامة ودراسات استرجاع المعلومات في بيئة الويب بصفة خاصة، سواء في البيئات الاصطناعية المعملية أو البيئات الحقيقية التشغيلية.

الفصل الأول

تمثيل المعرفة
واسترجاع المعلومات:
نظرة عامة

◄ 1 مقدمة

يرجع تاريخ نظم تمثيل المعرفة واسترجاع المعلومات إلى بدايات النصف الثاني من القرن التاسع عشر، وبالتحديد إلى عام 1876 عندما وضع ميلفل ديوي[1] Melvil Dewey أسس تمثيل المعرفة الحديث من خلال خطة التصنيف المعروفة باسمه كأداة أساسية لتنظيم وإتاحة المعرفة (Wynar & Taylor, 1985). مع ذلك فإن مجال تمثيل المعرفة واسترجاع المعلومات لم يصبح مجالاً محورياً للبحث ضمن مجالات علم المعلومات إلا مع نهاية الحرب العالمية الثانية. ومنذ ذلك التاريخ بدأت جهود مكثفة لتطوير هذا المجال الخصب، حيث جذب اهتمام الباحثين في مجالات متعددة. ويرجع ذلك بصفة أساسية إلى توظيف تكنولوجيا المعلومات منذ البداية في البحوث والتطوير بهذا المجال بدرجات متنوعة من التعقيد والنضج الأكاديمي.

يعد مصطلحا تمثيل المعرفة واسترجاع المعلومات المستخدمان في هذا السياق تطوراً للعديد من المصطلحات التي ظهرت منذ بداية القرن العشرين وحتى الآن، ومنها مصطلحات مثل التكشيف والاستخلاص، استرجاع المعلومات ومعالجة وتنظيم المعلومات، إدارة المعرفة.. إلخ.

وسيتم فيما يلي استعراض التطور التاريخي لمجال تمثيل المعرفة واسترجاع المعلومات مع التركيز على الملامح الأساسية التي شهدتها كل فترة.

(1) ميلفل ديوي yeweD «livleM» htussoK siuoL ellivleM (10 ديسمبر 1851 – 26 ديسمبر 1931) بمدينة نيويورك وهو مطوِّر ومؤسس أشهر خطط التصنيف الحديثة والمعروفة باسمه (خطة تصنيف ديوي العشري (metsys noitacfiissalC lamiceD yeweD).

كما سيتم شرح المفاهيم الأساسية المستخدمة في سياق مجال تمثيل المعرفة واسترجاع المعلومات، ثم مناقشة المكونات الأساسية لنظم تمثيل المعرفة واسترجاع المعلومات Knowledge Representation and Information Retrieval، وسينتهي هذا الفصل بشرح وتوضيح المشكلة الأساسية التي يعالجها هذا المجال والتي يمكن إيجازها في كيفية الحصول على المعلومات الملائمة التي تلبي الاحتياجات المعرفية لمستفيد بعينه في الوقت المناسب. ونظراً للعلاقة الوثيقة بين مجال تمثيل المعرفة واسترجاع المعلومات، كما سنوضح فيما بعد، سنختصر المصطلح المستخدم في هذا الكتاب إلى تمثيل واسترجاع المعلومات؛ نظراً لأن المعرفة مفهومة ضمناً أنها الهدف الأساس من كل عمليات تجميع البيانات وتجهيزها ومعالجتها وإنتاج المعلومات وتنظيمها وإتاحتها.

1.1 مراحل تطور تمثيل المعرفة ونظم استرجاع المعلومات

إن تاريخ نظم تمثيل واسترجاع المعلومات ليس طويلاً، ومع ذلك فقد شهد تطوراً سريعاً خلال الربع الأخير من القرن العشرين، والذي يُنظر إليه على أنه مرحلة إزالة الغموض عن هذا المجال. ويرى الباحثون أن مجال استرجاع المعلومات مرّ بأربع مراحل أساسية بداية من مرحلة زيادة الطلب على المعلومات حتى مرحلة عصر المشابكة Networked Era الذي نعيشه حالياً. ونستعرض فيما يلي مراحل تطور نظم تمثيل واسترجاع المعلومات.

1.1.1 مرحلة زيادة الطلب (بداية الأربعينات إلى بداية الخمسينات)

أدت الحرب العالمية الثانية إلى سرعة وتيرة التطوير في مجالات العلوم والتكنولوجيا، والتي أسهمت بصورة كبيرة في ظهور مجال تمثيل واسترجاع المعلومات، حيث أدت الحرب إلى إنتاج عدد كبير ومذهل من الوثائق والتقارير الفنية التي تسجل نتائج أنشطة البحوث والتطوير في مجال الصناعة وخاصة في مجالات صناعة الأسلحة وإدارة العمليات. وقد أدى هذا الكم الهائل من الوثائق إلى

الحاجة إلى أساليب جديدة لمعالجة الوثائق للوصول إلى ما تتضمنه من معلومات، حيث إن البشرية لم تواجه من قبل هذه المهمة الصعبة، والتي تمثلت في التعامل مع هذا الكم الهائل من الوثائق المهمة دون النظر إلى الجوانب الأخرى الخاصة بمعالجة وإدارة المعلومات مثل الاختيار والبث والحفظ.

وقد أوضح فانفر بوش (Bush,1945, p101) أن أحد أهم نتائج الحرب العالمية الأولى زيادة الاهتمام بأنشطة البحث والتطوير التعرف إلى ما تتضمنه الوثائق التي نتجت عن تلك الحرب من معلومات. فقد أتاحت الحرب الوصول إلى كم كبير من نتائج البحوث السرية التي احتاجت إلى الدراسة والتحليل، ما يعد مؤشراً قوياً إلى أن البشرية دخلت في مرحلة التعمق والتوسع في التخصصات العلمية. وقد واجه المكشفون مشكلات كبيرة نظراً للحاجة إلى استيعاب هذا الكم الهائل من الوثائق واستخلاص النتائج، ويبدو أنهم لم يستطيعوا التعامل إلا مع قدر قليل ومحدود جداً من المعلومات بسبب عقم أساليب الوصول إلى المعلومات في ذلك الوقت. وقد أصبح من الواضح أنه توجد حاجة حقيقية وضغط شديد نحو أساليب أكثر كفاءة لتمثيل وتنظيم هذا الكم الهائل من المعلومات وخاصة في مجالات الكيمياء والبيولوجيا والصناعة.

ويمكن تلخيص أهمية وجود آليات لاسترجاع المعلومات في ما يلي: على سبيل المثال، تقوم دور النشر والطبع في مجال الكيمياء الحيوية بنشر نحو مليوني وثيقة سنوياً (Hiemstra, 2009). ما يشير إلى مدى صعوبة التعامل مع تلك الوثائق باستخدام الأساليب التقليدية في الوصول إلى المعلومات. وتشير الإحصاءات إلى أن الباحث الواحد يحتاج إلى ساعة على الأقل لقراءة بحثين، فإذا افترضنا جدلاً أن هذا الباحث يستطيع قراءة بحوث بـ 70 لغة مختلفة، وأنه يستطيع الوصول إلى كل الوثائق المنتجة في مجال الكيمياء الحيوية (مليونا وثيقة سنوياً) في حوزته وبين يديه ويمكنه قراءة دورية واحدة في اليوم وأن العام به 365 يوماً، فإنه بحاجة إلى 27.4 قرناً لقراءة مخرجات البحوث في عام واحد فقط في مجال الكيمياء الحيوية (Borko & Bernier,1975, P.6).

وعلى الرغم من أن عدد التقارير الفنية التي تم إنتاجها خلال فترة الأربعينات

والخمسينات لا يمكن تحديدها بدقة، حيث إن حجم هذه الوثائق يمكن تقديرها وفقاً للتقدير السابق لمجال الكمياء الحيوية، ومع هذا الكم الهائل من الوثائق لا يمكن للإنسان أن يعتمد حصرياً على مهاراته وذاكرته وملفاته الخاصة لتنظيم واسترجاع المعلومات بطريقة فعالة لتلبية الاحتياجات في تلك الظروف. وبناء على ذلك ظهرت الحاجة إلى جهود مكثفة في مجال تمثيل واسترجاع المعلومات، وقد تنتج عنها أيضاً الحاجة إلى تطوير نظم لأغراض استرجاع المعلومات على الرغم من أنها كانت نظماً يدوية مثل كشافات الربط المسبق التي تم تطويرها في البداية عام 1951 والتي كانت أدوات فعالة في ذلك الوقت (Swanson,1988).

1.1.2 النمو المتسارع (الخمسينات حتى الثمانينات)

تُعد هذه الفترة هي الفترة الذهبية في نمو وتطور مجال تمثيل واسترجاع المعلومات؛ حيث شهدت دخول واستخدام الحاسب الآلي في هذا المجال خلال الفترة من 1957ـ 1959، وذلك عندما استخدم هانز بيتر لوهان Hans Peter Luhn البطاقات المثقبة في معالجة ومضاهاة الكلمات المفتاحية وترتيب المواد إلى جانب الأعمال الفكرية المرتبطة بتحليل محتوى النصوص (Salton,1987). وقد أدى ظهور نظم الاسترجاع على الخط المباشر مثل ديالوج DIALOG في الستينات والسبعينات من القرن الماضي إلى الانتقال من نظم استرجاع المعلومات اليدوية إلى النظم المتاحة على الخط المباشر. وقد وصف هاهن (Hahn ,1996) النظم الرائدة التي تم تطويرها في هذه المرحلة بما يلي:

اشتملت هذه النظم على مجموعة مهمة من الملامح المتطورة مثل المكانز المتاحة على الخط المباشر، فرز النتائج، الدمج الآلي للمترادفات أثناء إجراء البحث، المنطق البوليني، البتر من جهة اليسار وجهة اليمين left and right hand truncation، البحث في المصادر المستهدفة، البحث باللغة الطبيعية في النصوص الحرة. كما أتاحت بعض النظم إمكانيات التجميع الآلي للبيانات، برامج لمراقبة معدلات الاستخدام، ومدى رضا المستفيدين عن النظم.

وقد أسهم في نمو ونضج نظم استرجاع المعلومات على الخط المباشر إلى

جانب تطوير الأساليب الآلية والتجارب التي تمت في مجال استرجاع المعلومات، التطورات التي تمت في تكنولوجيا الحاسبات في ذلك الوقت. وقد كرّس الباحثون في العديد من المجالات وخاصة علوم الحاسب مجهوداتهم للبحوث والتطوير في هذا المجال، وعلى الرغم من ذلك ظلت العديد من المشكلات الإضافية التي تحتاج إلى جهود بحثية مكثفة، حيث لخص سالتون (Salton,1987) في أحد كتبه هذه المشكلات بما يلي:

على الرغم من التقدم الكبير الذي حدث خلال الثلاثين عاماً الماضية في مجال معالجة النصوص واسترجاع المعلومات وخاصة في مجال تحرير النصوص وإنتاج الوثائق وتحديد كلمات الكشاف والتجميع الآلي وبناء الاستفسارات وبحثها آلياً؛ إلا أنه توجد حاجة إلى جهود مكثفة في مجال فهم النصوص Text Understanding والمعالجة الدلالية للمعلومات Informtion Syemantic Processing. من ثم فإن هذه المرحلة ركزت على توظيف إمكانيات الحاسبات الآلية في تمثيل واسترجاع النصوص، ولكن ظلت عمليات فهم وتحليل دلالات النصوص تمثل مشكلة كبيرة للباحثين.

1.1.3 مرحلة إزالة الغموض 1980 – 1990

على الرغم من وصف نظم استرجاع المعلومات سابقاً بأنها نظم تم تطويرها لخدمة الاحتياجات المتنوعة والمتغيرة للمستفيدين منها؛ إلا أن هذه النظم لم يتم تصميمها بحيث يمكن للمستفيد أن يبحث فيها مباشرة دون الحاجة إلى تدريب أو تقديم الدعم من جانب أخصائي المعلومات. بمعنى آخر أن أخصائيي المكتبات والمعلومات كانوا يقومون بإجراء البحث نيابة عن المستفيدين فيما عرف بوسطاء البحث Search Mediators، إضافة إلى أن عملية البحث باستخدام هذه النظم كانت مكلفة للغاية، لما تتضمنه من مجموعة متنوعة من الرسوم، منها على سبيل المثال كُلفة تجهيزات الاتصال عن بعد Telecommunication، كُلفة الاتصال نفسه، رسوم اشتراكات قواعد البيانات.. إلخ، كما أن الرسوم كان يتم تحصيلها مقابل كل عملية بحث تتم. ومن ثم فمصطلح المستفيد النهائي End users الذي استُخدم للإشارة إلى

أصحاب الاحتياجات المعرفية لم يكن يمثلهم تمثيلاً حقيقياً، حيث إنهم لم يكونوا قادرين على إجراء البحث في تلك النظم بأنفسهم.

ومع الوقت بدأ مفهوم المستفيد النهائي يتغير تدريجياً مع ظهور الحاسبات الشخصية واستخدامها في عمليات البحث بنظم استرجاع المعلومات، و أيضاً مع بدايات تطبيق نظم الاسترجاع على الأقراص المدمجة CD- ROM والفهارس العامة المتاحة على الخط المباشر في منتصف الثمانينات من القرن الماضي.

وتجدر الإشارة إلى أنه في الماضي كانت نظم استرجاع المعلومات يتم إتاحتها من خلال نظم متنوعة مثل الحاسبات الآلية، طابعات النهايات الطرفية Printer Terminals، نظم البطاقات المثقبة الضوئية والميكانيكية.. الخ. وجدير بالذكر أن عملية التفاعل بين الباحث وتلك النظم لم تكن سلعة محفزة ولم تكن أيضاً سهلة للمستفيد User Freindly. وعندما تم استخدام الحاسبات الشخصية في استرجاع المعلومات وجد المستفيدون أنها أقل إزعاجاً وصعوبة من الأنظمة السابقة، نظراً لاعتمادها على حوارات فعلية للمستفيد مع الأجهزة، فيما عرف بالتفاعل بين المستفيدين والنظم.

لذلك ظهر فرع جديد من فروع علم المعلومات اهتم بالسلوك المعلوماتي للإنسان Information seeking Behavior وركز على تفاعل الإنسان مع الحاسبات Human Computer Interaction. وقد ساعد ظهور نظم الأقراص المدمجة والفهارس العامة المتاحة على الخط المباشر Online Public Access Catalogs – OPACs على إزالة الغموض وفض الالتباس الذي كان يكتنف عمليات البحث في تلك النظم وأصبح المستفيد قادراً على إجراء عملية البحث بنفسه، ولم يعد المستفيد يتأثر بكُلفة الاتصال عند إجراء البحث على الأقراص المدمجة ونظم الفهارس المتاحة على الخط المباشر. ومنذ ذلك الوقت أصبحت نظم استرجاع المعلومات أنظمة تم تطويرها لاستخدامها من جانب المستفيد النهائي، ما أثر بصورة كبيرة في انتشار تلك النظم وتطويرها نظراً للتفاعل الدائم من جانب المستفيد معها.

1.1.4 عصر الشبكات (التسعينات حتى الآن)

كانت نظم استرجاع المعلومات ــ حتى بداية التسعينات ــ نشاطاً مركزياً؛ حيث يتم إدارة قواعد البيانات التي تُعد المكون الأساسي لأي نظام استرجاع معلومات من خلال مقر مركزي واحد. فإذا كان الناس بحاجة إلى البحث في أكثر من نظام استرجاع معلومات، فعليهم أن يقوموا بالاتصال بكل قاعدة بيانات على حدة. ومع ظهور شبكات المعلومات وانتشار استخدامها ظهرت أنماط جديدة من البحث أطلق عليها البحث الموزع Distributed Searching الذي يسمح للمستفيدين بدخول قواعد البيانات والبحث فيها دفعة واحدة باستخدام البنية التحتية لشبكات المعلومات. ومن ثم لم تعد نظم استرجاع المعلومات قاصرة على نظام مركزي في موقع جغرافي واحد. وقد ساعد تقدم الإنترنت وتطوير إمكانيات الاتصال بها على تحويل هذا الأمر إلى حقيقة من خلال توفير البنية التحتية للاتصال البيني بين الشبكات المتنوعة والموزعة على مناطق جغرافية متعددة. فإلى جانب الملامح الخاصة بالبحث الموزع، أعادت الإنترنت صياغة مجال استرجاع المعلومات، ويسرت التعامل مع أساليب جديدة لمعالجة المعلومات، منها الطرق الإحصائية. فلم يسبق في التاريخ أن تم استخدام أو تطبيق النظم الإحصائية لمعالجة الكلمات المفتاحية مع هذا الكم الهائل من الروابط الفائقة ذات البنيات المتماسكة ومعلومات الوسائط المتعددة، كما لم يسبق في التاريخ أن قام هذا العدد الهائل من المستفيدين من إجراء البحث بنظم استرجاع المعلومات دون الحاجة إلى وسطاء أو مساعدة من أخصائيي المكتبات والمعلومات. وكنتيجة لذلك فإن جودة عملية تمثيل وتنظيم واسترجاع المعلومات في هذه البيئة تعقدت كثيراً، ما دعا إلى ظهور مصطلح جديد وهو مصطلح تنظيم الفوضى Organizing Chaos لوصف الوضع الذي ظهر مع بدايات انتشار الإنترنت وعلى وجه الخصوص محركات بحث الويب Web Search Engines. لذلك فقد أصبح استرجاع النصوص الكاملة Full Text Retrievel هو النمط السائد وليس الاستثناء في الاسترجاع على الإنترنت، كما ساعدت الإنترنت على سرعة انتشار تقنيات استرجاع المعلومات التي كان يتم اختبارها مسبقاً في المعامل، بحيث انتشرت نظم استرجاع معلومات التي تعمل على الإنترنت، ولعل أبرزها محركات بحث الويب مثل Google, Yahoo,

Bing, Ask Jeeves، وعموماً فإن نتائج البحوث الخاصة ببيئة المختبرات يتم تطبيقها بصورة موسعة في نظم تمثيل واسترجاع المعلومات على الإنترنت.

وعلى الرغم من أن المرحلة الرابعة وهي مرحلة محركات بحث الويب قد أثرت في كل أنماط العمل بقواعد البيانات ونظم استرجاع المعلومات التقليدية وفي سلوكيات المستفيدين، إلا أن هذه المرحلة نفسها مرت بالعديد من المتغيرات وبدأت تركز في السنوات الاخيرة على تطبيقات الذكاء الاصطناعي والويب الدلالي في عمليات التمثيل والبحث والاسترجاع التي سيتم تناولها بالتفصيل عند التعرض لتاريخ محركات البحث.

1.2 مفاهيم أساسية

يهتم هذا الكتاب بأربعة مفاهيم أساسية هي: هرم المعلومات، تمثيل المعرفة، استرجاع المعلومات، والعصر الرقمي. ويحظى كل مفهوم من هذه المفاهيم بمجموعة من المترادفات التي يمكن تفسيرها أو فهمها بطرق مختلفة وفي سياقات متنوعة. وسيتم فيما يلي توضيح هذه المفاهيم المختلفة التي يتضمنها هذا الكتاب.

1.2.1 هرم المعرفة

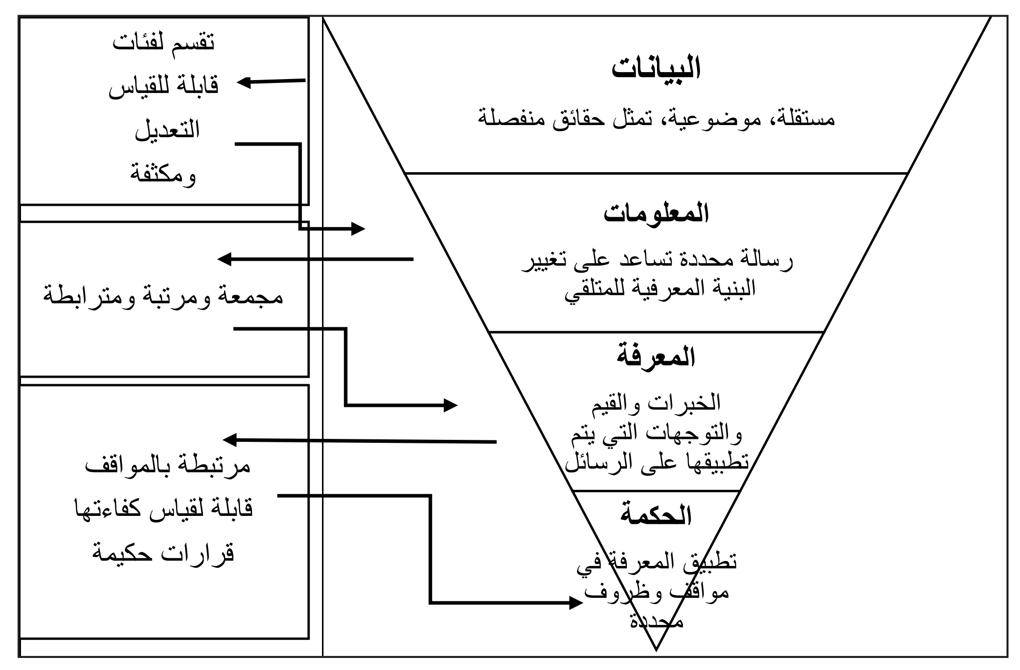
اهتم العديد من الباحثين بتفسير هرم المعلومات وتمييز عناصره التي تشمل البيانات، والمعلومات والمعرفة، والحكمة (Meadow,1992)، ويجب في هذا السياق تمييز مكونات هرم المعلومات وما يتضمنه من عناصر، وعلاقة كل مصطلح فيه بباقي المصطلحات. ويوضح الشكل التالي مكونات هرم المعلومات بعناصره الأربعة:

- **البيانات Data**: هي مجموعة من الحقائق الموضوعية الخام غير المترابطة وغير المنظمة. ويمكن لهذه البيانات أن تكون كمية أو كيفية (إحصاءات، أرقام، وقائع، بيانات ببليوغرافية). وعادة ما يشار إلى البيانات بأنها المادة الخام للمعلومات، حيث تتحول البيانات إلى معلومات عندما يتم تجميعها وتنظيمها وتصنيفها وتنقيحها وتحليلها ووضعها في إطار واضح ومفهوم

للمتلقي. فالبيانات الببليوغرافية لكتاب تشمل المؤلف والعنوان وبيانات النشر وبيانات الوصف المادي.. إلخ، وبيانات الشخص تشمل اسمه وعنوانه وتاريخ ميلاده ورقمه القومي ورقم جواز السفر وحالته الاجتماعية. ويتم تجميع تلك البيانات في بطاقات للتحقق من هوية الكيان (الكتاب أو الشخص) في صورة تسجيلات «تمثيل بيانات هذا الكيان».

> **«لاحظ استخدام مصطلح تمثيل البيانات في صورة تسجيلات وهو ما يتطابق مع عمليات معالجة البيانات والمعلومات، والتي تنطوي على عملية تمثيل للمحتوى وتجهيزه لعمليات البحث والاسترجاع الذي هو محور اهتمام هذا الكتاب».**

- **المعلومات:** تعرف عادة بأنها البيانات التي تمت معالجتها بحيث أصبحت مرتبطة بسياق معين ودلالات محددة. فالمعلومات هي بيانات توضع في إطار ومحتوى واضح ومحدد يساعد استخدامها على اتخاذ قرار في مواقف



شكل (1.1) مكونات هرم المعرفة

معينة. ويمكن التعبير عن المعلومات بأكثر من شكل منها النصوص المكتوبة، المسموعة، المرئية، المرسومة.. الخ. وعادة ما ينظر إلى المعلومات على أنها المحرك الأساسي لإحداث التغيير في البنية المعرفية للمتلقي. فبيانات الشخص لا يمكن من خلالها التعرف إليه، لكن يمكن تمييزه بوضوح من خلال بطاقة الهوية، جواز السفر، تسجيلاته الاستنادية التي تشتمل على بيانات تجميعية عن الكيان المطلوب تمييزه.

ونظراً لأن الكتاب يركز على موضوع استرجاع المعلومات فيجب تمييز المقصود بالمعلومات في هذا السياق. فقد تم استخدام مصطلحات مثل المعلومات والنصوص Texts والوثائق Documents بطريقة تبادلية في مجال استرجاع المعلومات. فالوثائق يمكن تصنيفها وفقاً لسعرها والذي يمكن من خلاله وضعها في أعداد وإحصاءات، والذي يعد المكوّن الأساسي لإحصاءات المواد بمؤسسات المعرفة، ومعظم هذه الوثائق تستغل مساحات، ويمكن أن يتم تدميرها أو أن تتعرض للتلف مع الوقت، إضافة إلى ذلك فإن الوثائق من الممكن أن تتضمن وسائط متعددة، فإذا كانت النصوص تشير إلى المعلومات النصية فقط، فإن الوثائق من الممكن أن تتضمن معلومات من وسائط متعددة (مزيج من المواد السمعية والبصرية والصور إلى جانب المعلومات النصية). من ثم فمن الواضح أن المعلومات تشتمل على كل من النصوص والوثائق والتي لها دلالة أوسع من الثلاثة مفاهيم (المعلومات، النصوص، الوثائق). وقد بدأ الاهتمام في السنوات الأخيرة بإجراء بحوث ودراسات عن الاسترجاع من الفقرات Passage Retrieval في مقابلة استرجاع الوثائق (Sparck Jones, 2000) ويهتم استرجاع الفقرات والذي يطلق عليه أيضاً في بعض الأحيان استرجاع المعلومات، بإيجاد المعلومات ذاتها أو الفقرات نفسها (مثل فقرات أو أجزاء محددة من الوثيقة) التي يحتاج إليها المستفيد. ويركز استرجاع الوثائق على الوثيقة كاملة للمستفيد النهائي حتى لو كان المستفيد لا يحتاج منها إلا إلى جزء أو فقرة صغيرة. من ثم فمصطلح معلومات في هذا السياق يشير إلى مفهوم شامل لمعالجة كافة أشكال وأنواع مواد وحاويات المعلومات سواء كانت نصية أو غير نصية بما في ذلك الكيانات بأكملها مثل الكتب والمقالات أو أجزائها مثل الملخصات والفقرات.

- **المعرفة**: هي المعلومات التي تم فهمها وتحليلها واستيعابها واستعمالها لإنجاز فعل معين أو اتخاذ قرار في ظروف معينة. فالمعرفة لا تقتصر على الأشياء الظاهرة والملموسة مثل القرارت بل تشمل أيضاً المهارات والخبرات الشخصية والتفسيرات والتحليلات والاستنتاجات التي يضيفها الأفراد والجماعات، والتي يتم من خلالها اتخاذ القرارت. ويتم تحصيل المعرفة من المعلومات المتاحة للشخص من مصادر المعلومات التي يتم الوصول إليها من خلال أدوات تنظيم وإتاحة المعلومات.

وتجدر الإشارة إلى أن المعرفة هي مجموع ما يمتكله الفرد من مقومات تمكنه من أداء مهام وإنجاز أعمال وحل مشكلات. كما أنها رأس المال البشري الذي تمتكله المجتمعات، فمجتمعات المعرفة هي المجتمعات التي تمتلك رأس مال بشرياً قادراً على أداء مهام وإنجاز أعمال وابتكار حلول لمشكلات الحياة اليومية، بحيث يمكنها تصدير تلك الحلول في صورة تطبيقات وإرشادات. فعلى سبيل المثال، الطبيب الذي يمتلك المعرفة هو رأس مال بشري يستطيع حل مشكلات صحية للعديد من المرضى، المبرمج الجيد هو رأس مال بشري يمتلك المعرفة التي تمكنه من بناء تطبيقات تحقق رفاهية المجتمعات. فإذا نظرنا إلى أهم شركة تأجير سيارات في العالم، وهي «أوبر» على سبيل المثال، نجد أنها لا تمتلك أي سيارة، وإنما تمتلك تطبيقاً لمعرفة ابتكرها رأس مال بشري استطاع توظيف البيانات والمعلومات المتاحة في بناء تطبيق مبتكر يحل مشكلة يواجهها الناس في حياتهم اليومية.

وقد حاول العلماء التمييز بين عناصر الهرم المعرفي من الناحية الرياضية بأساليب متنوعة، لعل أبرزها التعبير عن العلاقة بين البيانات والمعلومات والمعرفة بالمعادلة التالية:

$I = c + d$ **(I= المعلومة، d= البيانات، c = السياق)**

المعلومات تعادل كمّ البيانات التي يتم استخدامها في سياقات مختلفة.

كما عبر الباحثون عن العلاقة بين المعلومات والمعرفة بالمعادلة التالية

$K = I * U$ **(K = المعرفة، I = المعلومات، U = الاستعمال)**

المعرفة تعادل كم المعلومات مضروباً في عدد مرات استخدامها. وتجدر الإشارة إلى أن تحويل تلك المفاهيم إلى قياسات وطرق رياضية للحساب ليس بالأمر السهل؛ لأن كثيراً من تلك المفاهيم عادة ما يكون غير ملموس Intangible. ويمكن تخيل الأمر عند التعامل مع قاعدة بيانات تشتمل على مليون تسجيلة مثلاً، فحجم المعرفة الذي تتيحه هذا القاعدة للمستفيدين منها يعادل عدد التسجيلات المتاحة بها (مليون وحدة معلوماتية)، ونفترض أنه يتم استخدامها 1000 مرة يومياً وفي كل مرة يتم فحص 10 وحدات معلوماتية، بالتالي يكون حجم المعرفة التي توفرها تلك القاعدة يعادل عدد الوحدات المعلوماتية المستخدمة في عدد مرات استخدامها (1000*10) يعادل 10,000 وحدة معرفية.

كما تم ابتكار العديد من الطرق لقياس المعرفة منها طريقة القياس التي وضعها البنك الدولي، والتي تعرف بمنهجية قياس المعرفة Knowledge Assessment Methodlogy – MAM والتي تعد مقياساً تفاعلياً تم تطويره ضمن برنامج المعرفة من أجل التنمية Knowledge for Development –K4D. ويشتمل المقياس على 148 متغيراً هيكلياً نوعياً structural and qualitative variables وذلك لعدد 146 دولة حول العالم لقياس أداء تلك الدول في 4 مقومات أساسية لقطاع اقتصاد المعرفة وهي كالتالي:

- نظام الحوافز الاقتصادية والمؤسسية Economic Incentive and Institutional Regime
- التعليم Education
- الابتكار Innovation
- تكنولوجيا المعلومات والاتصالات Information and Communications Technologies

ويتم حساب المتغيرات بنظام درجات من صفر إلى 10، ثم يتم معادلة درجات الدولة مقارنة بالدول الأخرى التي معها في نفس المجموعة. وتحدد منهجية قياس المعرفة KAM مؤشر اقتصاد المعرفة الكلي (Knowledge Economy Index (KEI

ومؤشر المعرفة Knowledge Index (KI) بكل دولة من الدول الـ 146. ويشير مؤشر المعرفة إلى قدرة الدولة على إنتاج ونشر المعرفة، في حين يشير مؤشر اقتصاد العرفة KEI إلى قدرة الدولة على توفير البيئة المحفزة للأعمال Prevailling Bussiness Environement والتي يتم فيها توفير المعرفة المحفزة للأنشطة الاقتصادية والتي تحقق التنمية والخير للمجتمع[1] World Bank (2006).

كما قامت مؤسسة محمد بن راشد آل مكتوم للمعرفة بإعداد مؤشر للمعرفة أطلق عليه مؤشر المعرفة العالمي، والذي يعد أكثر المقاييس ثباتاً واستمرارية في الصدور منذ عام 2015 حتى الآن. يُعنى مؤشر المعرفة العالمي بقياس المعرفة بمختلف أشكالها وتجليّاتها بهدف دعم جهود تحقيق التنمية المستدامة. وهو عبارة عن خلاصة جهد مجموعة من الخبراء والمتخصصين في مختلف المجالات مثل التعليم بمختلف مراحله وأنواعه والاقتصاد والبحث والتطوير والابتكار والتكنولوجيا وغيرها.

ويعتمد هذا المؤشر على تجميع تراكمي للبيانات والمعلومات من خلال استقصاءات تستند إلى بيانات موثوقة ومحدثة ومنهجية للمقارنة بين الدول التي يشملها المؤشر والتي تمت المقارنة بينها في 7 قطاعات رئيسة هي:

- التعليم قبل الجامعي
- التعليم التقني والتدريب المهني
- التعليم العالي
- البحث والتطوير والابتكار
- تكنولوجيا المعلومات والاتصالات
- الاقتصاد
- البيئات التمكينية

---

(1) World Bank (2006). Knowledge Assessment Methodology. «World Bank Institute.» World Bank, Washington, (33 p.). http://siteresources.worldbank.org/KFDLP/Resources/KAM_Paper_WP.pdf

والمشروع بكافة تفاصيله متاح على منصة المعرفة للجميع Knowledge4all والتي يمكن الوصول إليها من خلال الرابط التالي:

http://www.knowledge4all.com/ar/115/Pages

- **الحكمة**: تم تعريف الحكمة على أنّها حالة أو صفة تمكن الفرد من إصدار الأحكام المقبولة من جانب الآخرين، لأنها عادة ما تتسم بالبصيرة Insight والحكم العادل. والحكمة هي هبة إلهية غير مرتبطة بكم المعلومات والمعارف التي يملكها الفرد ولكنها مرتبطة ببصيرته ومدى صفائها. لذلك وصفها المولى عز وجل بأنها وحي وهبة تؤتى منه، كما في قوله عز وجل في الآيات التي وردت فيها الحكمة:

[يُؤْتِي الْحِكْمَةَ مَنْ يَشَاءُ، وَمَنْ يُؤْتَ الْحِكْمَةَ فَقَدْ أُوتِيَ خَيْراً كَثِيراً] ٢٦٩ البقرة، [ذَٰلِكَ مِمَّا أَوْحَىٰ إِلَيْكَ رَبُّكَ مِنَ الْحِكْمَةِ] ٣٩ الإسراء، [وَلَقَدْ آتَيْنَا لُقْمَانَ الْحِكْمَةَ أَنِ اشْكُرْ لِلَّهِ] ١٢ لقمان، [وَشَدَدْنَا مُلْكَهُ وَآتَيْنَاهُ الْحِكْمَةَ وَفَصْلَ الْخِطَابِ]٢٠ ص.

من ثم فالحكمة هي قمة هرم المعلومات، وتأتي بعد المعرفة ويتسم أصحابها بالقدرة على القيادة وإلهام وتعزيز الدوافع لدى الآخرين. لذا فمن أهم عناصر اختيار القيادات والمديرين في المؤسسات هو مدى تمتعهم بالحكمة التي تمكنهم من اتخاذ القرارت السليمة في المواقف وفي الوقت المناسب.

ونستكمل فيما يلي مجموعة المفاهيم الأساسية التي يتناولها هذا الكتاب لتحديد المفاهيم المقصودة والمعاني المستهدفة لتلك المفاهيم.

1.2.2 تمثيل المعلومات
Information Representation

أياً كان شكل المعلومات، توجد حاجة أساسية لتمثيل تلك المعلومات قبل أن تصبح قابلة للاسترجاع. ويقصد بتمثيل المعلومات هنا، اشتقاق مجموعة من البيانات (مثل العناوين والكلمات المفتاحية والعبارات.. إلخ) من الوثيقة أو تخصيص

مصطلحات (مثل الواصفات ورؤوس الموضوعات) للوثيقة، من ثم يمكن التعرف إلى مضمونها وتمييزها وتمثيلها. وعادة ما يتم أداء عملية تمثيل المعلومات من خلال مزيج من العمليات تشمل: الاستخلاص، التكشيف، التصنيف، التلخيص والاشتقاق.

وعلى الرغم من أن معالجة المعلومات Information Processing وإدارة المعلومات Information Management لهما معانٍ مختلفة عن بعضهما بعضاً، إلا أنهما أحياناً ما يتم استخدامهما كمرادفات لتمثيل المعلومات. فبينما تتم الإشارة إلى معالجة المعلومات على أنها طريقة التعامل مع المعلومات لأغراض الاسترجاع How information Is Handeled for Retrieval Purposes، تتعامل إدارة المعلومات مع مجال واسع من الأنشطة المرتبطة بالمعلومات تتراوح بين اختيار وحفظ المعلومات.

ويستخدم في هذا الكتاب مصطلح تمثيل المعلومات ليغطي الجوانب والطرق المختلفة لإعداد بدائل أو تمثيل الوثائق Document Surrogate or Representations مثل الكشافات والمستخلصات، وذلك لأغراض استرجاع المعلومات.

1.2.3 الحاجة والطلب والاسترجاع

يتم النظر إلى مجال طلب المعلومات على أنه مجال موضوعي واسع النطاق يغطي كلاً من جوانب التمثيل والاسترجاع (Sparck Jones & Willett,1997) ويتم الإشارة إلى البُعد الخاص بالاسترجاع على أنه إتاحة المعلومات Information



شكل رقم (1.2) مراحل عمليات إدارة المعلومات

Access أو طلب المعلومات Informationn Seeking ويمكن النظر إلى هذه المصطلحات على أنها مرادفات لمصطلح الاسترجاع. ذلك على الرغم من أن كلاً منها له توجه ضمني خاص به. فالمصطلح «إتاحة المعلومات» يركز على جوانب الحصول على المعلومات، بينما يهتم مصطلح طلب المعلومات بالجوانب الخاصة بالمستفيد الذي ينخرط في نشاط المعلومات، أما البحث عن المعلومات Information Searching فيركز على كل ما يتعلق بكيف يتم البحث عن المعلومات. علاوة على مجموعة المصطلحات السابقة، ظهرت في السنوات الأخيرة مجموعة من المصطلحات التي يتم تداولها واستخدامها بكثافة في مجال استرجاع المعلومات تشمل التنقيب عن البيانات Data Mining، واكتشاف المصادر Resources Discovery. وتجدر الإشارة إلى أن هذين المصطلحين عادة ما يستخدمان في مجال الأعمال التجارية وفي بيئة المشابكة، ومن المتوقع أن يصبحا من المصطلحات الثابتة التي يتم تداولها بين المتخصصين في مجال استرجاع المعلومات في المستقبل.

ومن المعاني الأخرى التي تستخدم للدلالة على مفهوم استرجاع المعلومات مصطلح تخزين المعلومات Information Storage، والذي يتعامل أساساً مع تسجيل وتخزين وحفظ المعلومات. ورغم ذلك، فإن هذا المفهوم قد أصبح تدريجياً ممارسة قديمة لمفهوم حفظ المعلومات، حيث لم يعد تخزين المعلومات أمراً مهماً نتيجة للتطورات التكنولوجية المتسارعة. وقد تطور هذا المفهوم وأصبح يستخدم بصورة أوسع للدلالة على طرق وأساليب خزن وإتاحة المعلومات.

1.2.4 العصر الرقمي Digital Age

عادة ما يتم التفرقة بين المصطلح «رقمي» في مقابل المصطلح «تناظري»، وكلا المصطلحين مرتبط باستخدام التكنولوجيا الإلكترونية. وقد قامت شركة تيك تارجت (Tech Target, 2001)، وهي إحدى الشركات التي تهتم بتعريف المصطلحات التكنولوجية، بتعريف التكنولوجيا الرقمية بأنها:

«أحد أنماط التكنولوجيا الإلكترونية التي تقوم بتجميع وتخزين ومعالجة البيانات في وضعين أساسيين هما موجب وغير موجب». ويتم تمثيل الموجب بالرقم 1 وغير الموجب بالرقم صفر. لذلك فإن البيانات التي يتم نقلها وتداولها في البيئة الرقمية يتم التعبير عنها بمجموعة من سلاسل الأصفار والآحاد. أما قبل ظهور التكنولوجيا الرقمية، فكان النقل الإلكتروني يقتصر على التكنولوجيا التناظرية والتي تنقل البيانات في صورة إشارات إلكترونية بترددات متفاوتة في السعة، والتي يتم تحميلها على حامل الموجات Waive Carrier بترددات محددة. ويُعد البث الإذاعي والتلفزيوني والتليفون من أبرز النماذج التقليدية للتكنولوجيا التناظرية. ومع تقدم الحاسبات وشبكة الإنترنت وغيرها من أنماط تكنولوجيا المعلومات دخل الإنسان في العصر الرقمي بصورة كبيرة. وقد تم العديد من أنشطة البحث والتطوير المرتبطة بمجال استرجاع المعلومات في تلك البيئة الرقمية.

1.3 مفاهيم مرتبطة بمجال استرجاع المعلومات

سيتم فيما يلي استعراض مجموعة من المفاهيم الأساسية ذات العلاقة الوثيقة بمجال استرجاع المعلومات وتشمل: قواعد البيانات، آليات البحث، اللغة، واجهات التعامل. ويُعد البشر (بمن فيهم المستفيدون، وأخصائيو المعلومات)، وعمليات المعالجة والنظم، ثلاثة مكونات متداخلة تعمل معاً في مجال تمثيل واسترجاع المعلومات في البيئة الرقمية التي تتأثر بقوة بهذه المكونات الثلاثة.

1.3.1 تنظيم المعلومات

هو وضع المعلومات في سياق يمكن من خلاله الوصول إليها عند الحاجة في أقل وقت وبأقل مجهود. والمقصود بالسياق هنا هو وضع آلية للتنظيم تيسر عمليات الإتاحة والوصول إلى المعلومات. وعادة ما يتم تمثيل المعلومات من خلال أدوات تساعد على تيسير تداولها يطلق عليها: مصادر المعلومات / مواد المعلومات / أوعية المعلومات / الإنتاج الفكري. وتشير كل هذه المصطلحات

إلى: الكتب / الدوريات / المخطوطات / الخرائط / الصور / المصغرات الفيلمية / ملفات الكمبيوتر / النوت الموسيقية / الوثائق / الرسائل الجامعية / الأشكال والنماذج / مواقع الويب.. إلخ.

- والغرض الأساسي من تنظيم المعلومات هو تيسير عمليات استرجاعها من خلال نظم استرجاع المعلومات والتي تشمل: الببليوغرافيات، الفهارس، أدوات الإيجاد، السجلات، المرافق الببليوغرافية، قواعد البيانات، أدلة الويب، محركات البحث، ما وراء المحركات، البوابات، أدوات الاكتشاف.. إلخ.

وتعمل كل أدوات تنظيم واسترجاع المعلومات على تيسير سبل الوصول إلى المعلومات لتحقيق الأهداف التالية:

- **إيجاد مصادر المعلومات**: يساعد على التحقق من أن المعلومات موجودة ومتاحة ويمكن الوصول إليها، مثل الحاجة إلى كتاب معين.

- **تمييز الأعمال بالمواد التجميعية**: يساعد على التحقق من أي جزء من الأعمال التجميعية موجود ومتاح ويمكن الوصول إليه (مثل الحاجة إلى مقالة بدورية).

- تجميع المواد معاً بصورة منتظمة يساعد على بناء **مستودعات بالوثائق المنظمة في المكتبات والأرشيفات والمتاحف وملفات الإنترنت** وغيرها من المستودعات.

- **تيسير عمليات الاستشهاد المرجعي**: بمصادر المعلومات وفقاً لقواعد معيارية.

- **تيسير سبل الإتاحة بنقاط إتاحة متنوعة**: مثل المؤلف والعنوان والموضوع وغيرها.

- **تيسير سبل تحديد مواقع وأماكن حفظ المواد** التي يوجد بها نسخ يمكن الوصول إليها.

ومن المعروف أنه توجد خمس طرق أساسية لتنظيم المعلومات وعادة ما يشار إليها بالمختصرة LATCH، والتي تمثل الموقع، والترتيب الهجائي، الزمني، الفئات، الهرمي:

1. **الموقع Location** ويستخدم في تنظيم المعلومات المتعلقة بالطرق والمدن والمواقع المهمة مثل الآثار والآبار والحفريات..الخ.
2. **الترتيب الهجائي Alphabet** ويستخدم في القواميس والموسوعات والكشافات وقوائم الأسماء وغيرها من المعلومات النصية..الخ.
3. **الوقت Time** يستخدم في ترتيب الأحداث التاريخية والجارية مثل المعارض والبرامج.. إلخ.
4. **الفئات Category** ويستخدم هذا النمط من الترتيب في تجميع الفئات المتشابهة كما هو الحال في تجميع المواد في فئات المواد بالمراكز التجارية والصيدليات ومواقع الويب. وقد يكون الترتيب وفقاً للنوع أو الشكل أو وفقاً للفئة العمرية.
5. **الترتيب الهرمي Hierarchy** يستخدم في عمليات التصنيف للمواد حسب علاقتها ببعضها بعضاً مثل التصنيف البيولوجي وتصنيف الموضوعات، وعادة ما يعتمد الترتيب الهرمي على وجود علاقة هرمية بين المواد، بحيث يتم تقسيمها من العام إلى الخاص.

1.3.2 استرجاع المعلومات

يشير مصطلح استرجاع المعلومات إلى أنه عملية بحث مجموعة من بدائل الوثائق، ويستخدم مصطلح وثيقة هنا على نطاق واسع لتحديد الوثائق التي تعالج موضوع معين. كما يتم الإشارة إليه على أنه أي نظام تم تصميمه لتيسير عملية بحث الإنتاج الفكري، ويطلق على هذا النظام مصطلح «نظام استرجاع المعلومات».

وعند تحديد مصطلح استرجاع المعلومات للدلالة على استرجاع الوثائق لابد من استبعاد الأنظمة التي لا تتعامل مع النصوص مثل نظم إدارة قواعد البيانات Database Management Systems ونظم الرد على الاستفسارات Questions Answering Systems. هذه النظم عادة ما يطلق عليها أنظمة استرجاع البيانات Data Retrieval Systems أو نظم استرجاع الحقائق Fact Retrieval Systems. وتتيح هذه الأنظمة استرجاع بيانات أو حقائق محددة تعبر عن معلومة محددة، وبعض هذه الأنظمة يتخطى مرحلة تقديم إجابات محددة إلى تقديم تحليل دقيق للنتائج في صورة أكثر ذكاءً، حيث تستخلص من البيانات المخزنة نتائج جديدة.

ومن الواضح أن مصطلح «استرجاع المعلومات» ليس مصطلحاً دقيقاً للدلالة على هذا النشاط الذي يتم تطبيقه فيه، حيث إن نظم استرجاع المعلومات لا تسترجع معلومات وإنما تسترجع بدائل لمصادر المعلومات. فمصطلح المعلومات يشير إلى شيء غير محسوس لا يمكن رؤيته أو سماعه أو الإحساس به، لأنه مرتبط بتغيير النمط المعرفي وتطوير البنية المعرفية للمتلقي، كما أن عملية الإعلام تتم عندما يحدث تغيير في البنية المعرفية للشخص في موضوع معين، من ثم إعطاء المستفيد وثيقة تتناول موضوعاً معيناً لا يعني إعلام المستفيد بالموضوع، وإنما الإعلام يحدث عندما يقوم المستفيد بقراءة الوثيقة وفهمها واستيعاب محتواها، ما يؤدي إلى إحداث تغيير في معرفته حول هذا الموضوع.

وعلى الرغم من أن المصطلح غير دقيق لوصف الموضوع، إلا أنه أكثر المصطلحات ملاءمة لأغراض مناقشة الموضوع بدقة، كما أنه المصطلح الذي استقر عليه الإنتاج الفكري المتخصص في الموضوع.

ومن الأنشطة الأساسية التي تقوم بها مؤسسات المعلومات، الإجابة عن الاستفسارات، والتي يمكن النظر إليها على أنها من أنشطة استرجاع المعلومات. وتسعى أنشطة الرد على الاستفسارات إلى توفير إجابات مباشرة عن استفسارات المستفيدين ومن أمثلة هذه الاستفسارات: ما هو ارتفاع جبل ما؟ ما درجة حرارة ذوبان مادة ما؟ ما عنوان.. ؟.

وتتم الإجابة عن مثل هذه الاستفسارات من خلال البحث في المصادر المرجعية

وتوفير إجابات مباشرة عن الاستفسارات بدلاً من إحالة المستفيد إلى وثيقة تجيب عن الاستفسار. ويطلق على هذه النوعية المتميزة من الخدمات مصطلح الخدمة المرجعية. تعد هذه النوعية من الخدمات المرحلة الثانية في أنشطة استرجاع المعلومات، حيث تتضمن المرحلة الأولى استخدام نظم استرجاع المعلومات على اختلاف أنواعها مثل فهارس المكتبات، الكشافات، قواعد البيانات، محركات البحث أو حتى كشاف نهاية الكتاب لتحديد الوثائق التي تجيب عن استفسار معين. ويتم في المرحلة الثانية استخلاص الإجابة من الوثائق التي تم تحديدها في المرحلة الأولى. وتجدر الإشارة إلى أنه قد تم تطوير العديد من نظم استرجاع الحقائق التي يتم البحث فيها من خلال توجيه استفسارات في صورة تساؤلات باستخدام اللغة الطبيعية، ونظراً للتعقيد الشديد في تصميم مثل هذه النظم فإن معظم النظم المتاحة حالياً مقصورة على نوعية معينة من المعارف ذات البنية المحددة مثل نتائج الاختبارات وتنسيق الجامعات أو أكواد الطرق السريعة. كما توجد نوعية أخرى من النظم التي تقدم إجابات أو استفسارات تتعلق بالمواد الفيزيائية أو الكيميائية أو المعادلات الرياضية.. إلخ. ويطلق على هذه النوعية من النظم نظم استرجاع البيانات، كما يمكن أن يشار إلى البيانات في هذه النظم بمصطلح بنوك البيانات، ومن أمثلة هذه البنوك: البيانات الإحصائية، بيانات مواد الطاقة.. إلخ. وقد حظيت هذه النوعية من بنوك البيانات باهتمام كبير في السنوات الأخيرة في ظل تضخم حجم البيانات من ثم ظهرت الحاجة إلى معالجة البيانات الضخمة Big Data والربط بين البيانات الضخمة Linked Big Data، إضافة إلى معالجتها بأساليب جديدة تشمل التنقيب عن البيانات Data Mining والمعالجات الدلالية للبيانات Semantic Data Analysis. وقد كان لكل هذه التطورات أثر كبير في نظم استرجاع المعلومات التي سعت نحو توفير آليات للتعامل مع تلك التطورات.

وهناك نوع ثالث من نظم استرجاع المعلومات يعتمد على تخزين وبحث النصوص الكاملة للوثائق، بحيث يستطيع استرجاع أجزاء من تلك الوثائق التي تضاهي استراتيجيات البحث المستخدم في التعبير عن احتياجات المستفيدين. وبهذا تعد نظم الإجابة عن الاستفسارات ونظم استرجاع البيانات ونظم استرجاع النصوص،

أمثلة لنظم استرجاع المعلومات على الرغم من أن الإجابة عن الاستفسارات ونظم استرجاع النصوص تسترجع معلومات مباشرة للإجابة عن استفسارات معينة، بينما نظم استرجاع المعلومات تسترجع بدائل للوثائق وليس الوثائق نفسها وتحيل المستفيد إلى النصوص الكاملة. لكن في ظل التطورات التي شهدتها أدوات البحث أصبحت نظم استرجاع المعلومات قادرة على استرجاع بدائل الوثائق والبحث في النصوص والرد على استفسارات المستفيدين في نفس الوقت. ولعل أبرز مثال على ذلك ما يقدمه محرك البحث غوغل الذي يدمج كل فئات البحث في صندوق واحد، كما يتيح إمكانية البحث في كل فئة على حدة.

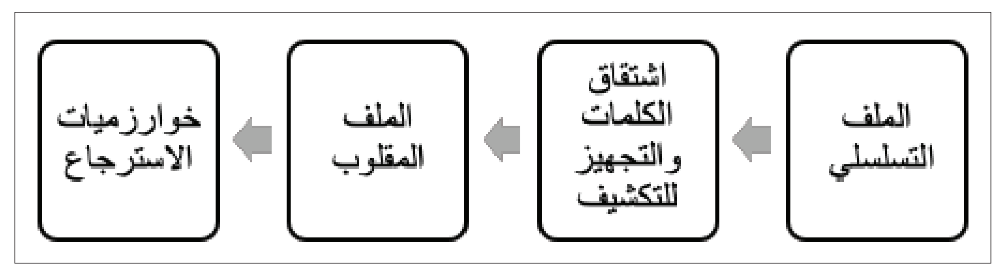
1.3.3 قواعد البيانات

تعد قواعد البيانات العمود الفقري وأحد المكونات الأساسية لنظم تمثيل واسترجاع المعلومات، حيث تشتمل على البيانات والمعلومات التي يتم تمثيلها وتنظيمها وفقاً لآليات عمل نظم استرجاع المعلومات التي سنتناولها بالتفصيل في هذا الكتاب. فالمفهوم التقليدي لقواعد البيانات التي تعرف بقواعد البيانات الببليوجرافية يشير إلى مجموعة من التسجيلات المتطابقة والتي يمكن تحليلها إلى حقول، والتي تُعد أصغر وأدق المكونات أو الوحدات التي تستخدم في عمليات البحث بنظم استرجاع المعلومات وفرز النتائج. ففي قاعدة بيانات الدوريات، على سبيل المثال، يوجد حقل يمثل بيانات التأليف وآخر يمثل عنوان المقالة.. الخ، وتستخدم هذه الحقول في عمليات البحث والتصفح والترتيب.

وتشتمل قواعد البيانات التقليدية على ملفين أساسيين هما الملف التسلسلي Sequential File والملف المقلوب Inverted File. ويُعد الملف التسلسلي مصدر قاعدة البيانات، حيث يشتمل على معلومات منظمة بنفس طريقة بنية الحقول والتسجيلات في قاعدة البيانات ويطلق عليه الملف التسلسلي، نظراً لأن التسجيلات مرتبة فيه ترتيباً تسلسلياً وفقاً لنفس تسلسل إدخالها بقاعدة البيانات.

أما الملف المقلوب، والذي يُعرف أيضاً بالملف الكشاف Index file، فيتيح

الوصول إلى الملف التسلسلي بناء على الصيغ البحثية ومدى تطابقها مع مصطلحات الكشاف المخزنة في الملف المقلوب. ويطلق عليه المقلوب نظراً لطريقة عرض المعلومات به حيث تأتي نقاط الإتاحة Access Point أولاً ثم المواضع Locators، وهو عكس الترتيب الذي توضع فيه المعلومات في الملف التسلسلي حيث تأتي المواضع أولاً ثم نقاط الإتاحة.



شكل (1.3) مكونات قاعدة البيانات

ويتضح من الشكل السابق أن قواعد البيانات تقوم بأربع عمليات أساسية لتجهيز الملفات لعمليات البحث والاسترجاع وهي:

- تجهيز الملف التسلسلي.
- بناء ملف الكشاف الذي يشتمل على الكلمات القابلة للتكشيف في كل تسجيلة.
- بناء الملف المقلوب الذي يشتمل على المصطلحات الكشفية ومواقعها بالتسجيلات.
- تطبيق خوارزميات الاسترجاع والتي تتضمن الوزن النسبي للمصطلحات الكشفية.

أما في النظم غير التقليدية مثل نظم الاسترجاع على الإنترنت، فإن قواعد البيانات تظل تشتمل على الملفات (التسلسلي والمقلوب)، إلا أن تركيب الملف التسلسلي على سبيل المثال قد يختلف عن تركيبه في النظم التقليدية على الخط المباشر؛ حيث إن التركيب في النظم غير التقليدية لا يأخذ شكل حقول وتسجيلات متطابقة في قواعد البيانات؛ فهو لا يتضمن حقولاً، وإنما يتم عرض المعلومات في شكل نثري، إضافة إلى أن المعلومات التي يتضمنها الملف التسلسلي ليست بدائل Surrogate للوثائق

جدول (1.1) نموذج لمكونات الملفات بقواعد البيانات

المحتوى	الكلمات المشتقة للتكشيف		الترتيب الهجائي	الوزن النسبي		
التسجيلات البيليوجرافية الكاملة	الكلمات المفتاحية	أرقام التسجيلات	أرقام التسجيلات	الكلمات	ارقام التسجيلات	الوزن
1	استرجاع	2، 4، 5	2، 4، 5	استرجاع	2	0.98
2	المعلومات	1، 2، 3	1، 2	استرجاع	4	0.70
3	نظم	1، 2، 5	1، 2، 3	استرجاع	5	0.85
4	معرفة	2، 3، 4	2، 3، 4	تمثيل	1	0.6
5	تمثيل	1، 2	1، 2، 5	تمثيل	2	0.84
				معلومات	1	0.66
				معلومات	2	0.75
				معلومات	3	0.85
				معرفة	2	0.55
				معرفة	3	0.64
				معرفة	4	0.90
				نظم	1	0.30
				نظم	2	0.67
				نظم	5	0.88

أو تسجيلات تلخص الوثائق، ولكنها جزء من محتوى الوثائق الأصلية المتاحة على الإنترنت، والتي يطلق عليها صفحات الويب. وفي نظم استرجاع المعلومات التقليدية، فإن الملفات التسلسلية تشتمل على بدائل للوثائق في صورة تسجيلات ببليوجرافية وصفية ومستخلصات أو ملخصات واشتقاقات لكلمات مفتاحية من بعض المواضع المهمة مثل العنوان، الملخص. كما أن المحتوى والتغطية اللذين تتضمنهما قاعدة البيانات يحددان المواد التي سيتم استرجاعها من النظام لكل عملية بحث.

1.3.4 آليات البحث
Search Mechanism

تتم عمليات البحث في قواعد البيانات من خلال توجيه استفسارات في صورة عبارات بحثية إلى محركات وأدوات البحث التي تقوم بدورها بتطبيق آليات البحث التي توفرها المحركات على الاستفسارات وتوجهها إلى قواعد البيانات لاسترجاع المعلومات التي يتم تمثيلها وتنظيمها بطرق ثابتة في ملفات قواعد البيانات، كما أوضحنا سابقاً. وتشتمل آليات البحث على إمكانيات متعددة من حيث مستوى التعقيد، والتي يتم تعريفها وتفسيرها وفقاً للخوارزميات Algorithms والإجراءات التي يتضمنها نظام استرجاع المعلومات. ويوجد بصفة عامة نموذجان أساسيان للبحث في محركات وأدوات البحث هما:

البحث الأساسي Basic search والبحث المتقدم Advanced search وتشتمل تقريباً معظم نظم استرجاع المعلومات على إمكانيات البحث البسيط والمتقدم، إلا أن إمكانيات البحث المتقدم تحتاج إلى مستفيد على كفاءة ووعي كاملين بإجراءات البحث وطرق صياغته؛ حيث إنها تقدم إمكانيات متنوعة ومتعددة في عمليات البحث كتلك التي يتم استخدامها أيضاً في الاختبارات المعملية لنظم استرجاع المعلومات. وفي السنوات الأخيرة اهتم العديد من نظم استرجاع المعلومات على الإنترنت بتطوير إمكانيات وآليات البحث المتقدم، لكي تتيح للمستفيد إمكانيات توجيه استفسارات معقدة لمحركات بحث الإنترنت.

وتشتمل إجراءات البحث على العديد من الإمكانيات التي توظفها نظم استرجاع

المعلومات في تحديد العلاقات بين الكلمات التي تشتمل عليها استفسارات المستفيدين منها الكلمات المفتاحية، البحث البوليني Boolean search الجذع Truncation التقارب Proximity.. الخ. ويحتاج المستفيد إلى مجموعة متنوعة من الخبرات والمهارات التي يحصل عليها من خلال التدريب حتى يتمكن من البحث بكفاءة وفعالية في نظم استرجاع المعلومات. أما النظم الحديثة والمتقدمة التي تشتمل على إجراءات بحث معقدة مثل البحث بالوزن Weighted Searching والتي يتم تصميمها خصيصاً لكي يتعامل معها فئات معينة تحصل على تدريب مكثف وتمتلك خبرات بحثية خاصة تلبي احتياجاتهم المعلوماتية والمعرفية المعقدة. وسوف يتم مناقشة هذه الآليات بصورة أكثر تفصيلاً في الفصل الحادي عشر.

1.3.5 اللغة Language

تُعد اللغة الوسيط الأساسي لنقل وتمثيل وعرض المعلومات سواء كانت مقروءة أو مكتوبة. وفي هذا السياق تُعد اللغة أحد المكونات الأساسية لتمثيل واسترجاع المعلومات. ويتم استخدام اللغة في إطار نظم تمثيل واسترجاع المعلومات بطريقتين أساسيتين هما: اللغة الطبيعية Natural Language واللغة المضبوطة أو المقيدة Controlled Vocabulary. فالطريقة التي يستخدمها المستفيدون في التعبير عن احتياجاتهم المعلوماتية في صورة استفسارات يُطلق عليها اللغة الطبيعية. أما في حالة استخدام لغة اصطناعية Artificial Language والتي تتضمن مصطلحات، تراكيب Syntax، ودلالات Semantics، يتم ضبطها وتقييدها من خلال قوائم مصطلحات محددة يطلق عليها اللغة المضبوطة أو المقيدة (Wellisch & Dowding, 1996).

ويوجد ثلاثة أنواع شائعة من اللغات المضبوطة هي: خطط التصنيف، وقوائم رؤوس الموضوعات والمكانز، ولكل منها استخدامه الخاص في نظم تمثيل واسترجاع المعلومات. وتتيح اللغة الطبيعية، بصفة عامة، قدرة كبيرة على التحديد والدقة والمرونة في تمثيل واسترجاع المعلومات، حيث لا يحتاج المستفيدون إلى التدريب عليها أو الممارسة لكي يتمكنوا من تطبيقها في عمليات البحث والاسترجاع، لأنها

الوسيلة الأساسية التي يستخدمونها في حياتهم اليومية للتواصل الشفاهي والمكتوب. وعلى العكس، فإن بناء وصيانة وتحديث اللغة المضبوطة تُعد أمراً مكلفاً، كما أن المستفيدين منها في حاجة إلى تعلم كيفية استخدامها والتدريب على ممارسة البحث واختيار المصطلحات من خلالها. ومع ذلك فإن اللغة المضبوطة تساعد على تقليص المشكلات والصعوبات التي توجد في اللغة الطبيعية مثل التعقيد، والغموض، وعدم الدقة في تمثيل واسترجاع المعلومات (Lansdale & Ormerod,1994). وتجدر الإشارة إلى أنه يوجد جدل كبير حول المقارنة بين اللغة الطبيعية في مقابل اللغة المضبوطة باسترجاع المعلومات يرجع تاريخه إلى نهايات القرن التاسع عشر، ومازال هذا الجدل قائماً حتى الآن. وتساعد اللغة المستخدمة في عملية التمثيل والاسترجاع، بدرجة كبيرة، على تحديد مستوى المرونة والحرفية أو التصنع في نظم استرجاع المعلومات. وسوف يتم مناقشة قضية اللغة في تمثيل واسترجاع المعلومات بشكل أكثر تفصيلاً في الفصل الخامس من هذا الكتاب.

◄ 1.3.6 واجهة التعامل Interface

ترى شاو (Shaw,1991) أن واجهة التعامل هي الجزء الذي يراه ويلمسه ويستمع إليه المستفيد عندما يتعامل مع أي نظام محوسب بصفة عامة، ونظم استرجاع المعلومات بصفة خاصة. ويشار إلى واجهات التعامل في إطار نظم تمثيل واسترجاع المعلومات بأنها التفاعل الذي يتم بين المستفيد والأنشطة التي يتعامل معها على النظام. كما أن هذا المكون يجعل المستفيد عنصراً واضحاً ومتداخلاً مع المكونات الثلاثة الأخرى لنظم تمثيل واسترجاع المعلومات (قواعد البيانات، آليات البحث، اللغة).

تُعد واجهة التعامل العنصر الحاسم في الحكم على مدى الصداقة للمستفيد User Friendly. فكما تم تحديدها بقانون مورز Moor's Law ؛ فالنظم الأكثر سهولة للمستفيد تجذب عدداً أكبر من المستفيدين من النظم المعادية للمستفيد User Hostile وفقاً لمعدلات الاستخدام، ويتم تحديد مدى كفاءة واجهة التعامل من خلال التفاعل معها وتقييم المعلومات التي تتضمنها مثل قوائم الاختيارات، أساليب العرض،

تصميم الشاشات، أنواع الخطوط وغيرها من الأمور المرتبطة بالقابلية للاستخدام Usability. وقد ركزت معظم النظم على استخدام التكنولوجيا المتأقلمة والفعالة Adaptive & Effective في تصميم وتنفيذ واجهات التعامل، أكثر من تركيزها على الجوانب البشرية لتمثيل واسترجاع المعلومات. من ثم تُعد واجهة التعامل العنصر المحدد لمدى نجاح أي نظام لتمثيل واسترجاع المعلومات، وخاصة إذا كان النظام يعمل في البيئة الرقمية.

بذلك يمكن القول بصفة عامة إن قاعدة البيانات بما تتضمنه من جداول وكشافات، آليات البحث، اللغة، وواجهة التعامل، هي مجموعة العناصر الجوهرية المكونة لأي نظام تمثيل واسترجاع معلومات، والتي يتفاعل معها المستفيد عند إجراء عمليات البحث والاسترجاع.

المصادر

- Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28(3), 289-299.
- Boole, G. (1854). An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities. Dover Publications.
- Borko, H., & Bernier, C. L. (1975). Abstracting Concepts and Methods. New York, Academic Press.
- Bush, V. (1945). As we may think. The Atlantic Monthly, 176(1), 101-108.
- Cleverdon, C. (1984). Optimizing Convenient Online Access to Bibliographic Databases. Information services and Use, 4, 37-47.
- Corbitt, kevin D.(1992). Calvin N. Moooers papers, 1930-1978 (CBI 81). Minneapolis: center for the history of computing. Charles babbage institute, university of Minnesota. retrieved October 3.2009, from www.libsci.sc.edu/bob/isp/mooers.htm
- Crouch, C., McGill, M., Lesk, M., Jones, K. S., Fox, E. A., Harman, D., & Kraft, D. H. (1996). Gerald Salton, March 8, 1927–August 28, 1995. Journal of the American Society for Information Science, 47(2), 108-115.
- Cuadra, C. A. (1964). Identifying key contributions to information science. American

Documentation, 15(4),289-295.

- Fischer, M. (1966). The KWIC index concept: A retrospective view. American Documentation, 17(2), 57-70.
- Garfield, E. (1997). A tribute to Calvin N. Mooers, a pioneer of information retrieval. The Scientist, 11(6), 9.
- Grosz, B. J., Jones, K. S., & Webber, B. L. (1986). Readings in natural language processing.
- Gull, C. D. (1956). Seven years of work on the organization of materials in the special library. AmericanDocumentation, 7(4),320-329.
- Gull, C. D. (1987). Information science and technology: From coordinate indexing to the global brain. Journal of the American Society for Information Science, 38(5), 338-366.
- Hahn, T. B. (1998). Pioneers of the online age. Historical studies in information science, 116-131.
- Harvey, john F. (1978). LUHN, Hans Peter (1896-1964). In Bohdan S.
- Henderson, madeline m. (1996). In memoriam: Calvin n. Mooers, October 24. 1919-december1.1994. journal of the American society for information science,47(9),659-661.
- Hiemstra, Djoerd. "Information retrieval models." Information Retrieval: searching in the 21st Century (2009): 2-19.
- Humphrey, S. M. (1992). Indexing biomedical documents: from thesaural to knowledge-based retrieval systems. Artificial Intelligence in Medicine, 4(5), 343-371.
- Koenig, M. E. (1987). The convergence of Moore's/Mooers' law's. Information processing & management, 23(6), 583-592.
- Lancaster, F. W. (1968). Information retrieval systems; characteristics, testing, and evaluation.
- Lansdale, M. W., & Ormerod, T. C. (1994). Understanding interfaces: a handbook of human-computer dialogue. Academic Press Professional, Inc..
- Larsen, P. S. (1999). Books and bytes: Preserving documents for posterity. Journal of the American Society for information science, 50(11), 1020-1027.
- Luhn, H. P. (1953). A new method of recording and searching information. American Documentation, 4(1),14-16.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.
- Luhn, H. P. (1961). Selective dissemination of new scientific information with the

aid of electronic processing equipment. American Documentation, 12(2), 131-138.

- McCandless, R. F. J., Skweir, E. A., & Gordon, M. (1964). Secondary Journals in Chemical and Biological Fields.Journal of Chemical Documentation,4 (3), 147-153.4
- Meadow, C. T., Boyce, B. R., & Kraft, D. H. (1992). Text information retrieval systems (Vol. 20). San Diego, CA: Academic Press.
- Mooers, C. N. (1960). Mooers' Law: Or, why some retrieval systems are used and others are not.American Documentation, 11(3), ii.
- Pao, M. L. (1989). Concepts of information retrieval. Englewood, Colo.: Libraries Unlimited.
- Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. Journal of the American Society for Information science, 27(3), 129-146.
- Robertson, S., & Tait, J. (2008). Karen Spärck Jones. Journal of the American Society for Information Science and Technology, 59(5), 852-854.
- Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. Journal of information science, 20(2), 108-118.
- Salton, G. (1987). The past thirty years in information retrieval. Journal of the American Society for Information Science, 38(5), 375-380.
- Saracevic, T. (1995). In memoriam: Gerard Salton (1927–1995). Information Processing and Management: an International Journal, 31(6), 787-788.
- Schultz, C. K. (1968). Luhn: Pioneer of Information Science.
- Shaw, Debora. (1991). The human- computer interface for information retrieval. Annualreview of information science and technology,26,155-195
- Shera, jesse h. (1978).taube, Mortimer (1910-1965). In bohdan s. wynar (ED.), Dictionary of American library biography (pp. 512-513).littleton, co: libraries unlimited
- Smith, E. S. (1993). On the shoulders of giants: From Boole to Shannon to Taube; The origins and development of computerized information from the mid-19th century to the present. Information Technology and Libraries, 12(2), 217.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.
- Sparck Jones, K.(ED.). (1981). Information retrival experiment. London: Butterworths.
- Sparck Jones, K.(1994).Finding the information wood in the natural language tree [Videotape]. Talk presented at the Grace Hopper celebration of women in computing meeting.41 min.
- Sparck Jones, K.(1995). Reflection on TREC.information processing & management,

31 (3), 291-314.

- Sparck Jones, K.(2000). Further reflections on TREC. information processing & management, 36 (1), 37-85.
- Sparck Jones, K.(2005). Some points in a time. computational linguistics, 31(1), 1-14.
- Sparck Jones, K.(2007).Automatic summarizing: the state of the art.
- Sparck Jones, K, and Willett, Peter (Eds.) (1997). Readings in information retrieval. San Francisco: Morgan kaufmann.
- Stevens, mary Elizabeth. (1968). H. P. luhn, information scientist. In Claire K. Schultz (ED.), H.P. Luhn: pioneer of information science: selected works (pp. 24-30). New York: Spartan Books.
- Swanson, D. R. (1977). Information retrieval as a trial-and-error process. The Library Quarterly, 47(2),128-148.
- Swanson, D. R. (1988). Historical note: Information retrieval and the future of an illusion. Journal of the American Society for Information Science, 39(2), 92-98.
- Swanson, Don R. "Historical note: Information retrieval and the future of an illusion."Journal of the American Society for Information Science39.2 (1988): 92-98.
- Tait, J. I. (2005). Natural Language Processing and Information Retrieval.
- Teach Target. (2001). Digital. Retrieved, December 3, 2008, From whatis. Techtarget. com
- Wellisch, Hans, and Martin Dowding. "[Indexing from A to Z." Journal of Scholarly Publishing, 28.1
- Wilks, Y. (2007). Karen Spärck Jones (1935-2007)[In Memoriam]. IEEE Intelligent Systems, 22(3), 8-9.
- Willett, P., & Robertson, S. (2007). In memoriam: Karen Spärck Jones. Journal of Documentation, 63(5).
- Wynar, B. S., Taylor, A. G., & Osborn, J. (1985). Introduction to cataloging and classification. Littleton, CO: Libraries Unlimited.

الفصل الثاني

مشكلة التمثيل

واسترجاع المعلومات

2 مقدمة

يستعرض هذا الفصل المشكلة الرئيسة التي تحاول كل أنظمة استرجاع المعلومات توفير حلول لها، سواء كانت هذه الحلول في البيئة الورقية أو الإلكترونية أو الرقمية. وتتمثل تلك المشكلة في جانبين أساسيين هما: الجانب الرياضي المتعلق بكفاءة النظام وقدرته على استرجاع كل الوثائق الصالحة والمقاييس المستخدمة في الحكم على الكفاءة وطريقة تطبيقها؛ والجانب الإجرائي المتعلق بإجراءات التمثيل والبحث بقواعد البيانات أو محركات البحث. ثم يستعرض الفصل تمثيل المعلومات والتحديات المتعلقة بعمليات التمثيل وآليات التغلب عليها.

2.1 المشكلة الأساسية لتمثيل واسترجاع المعلومات

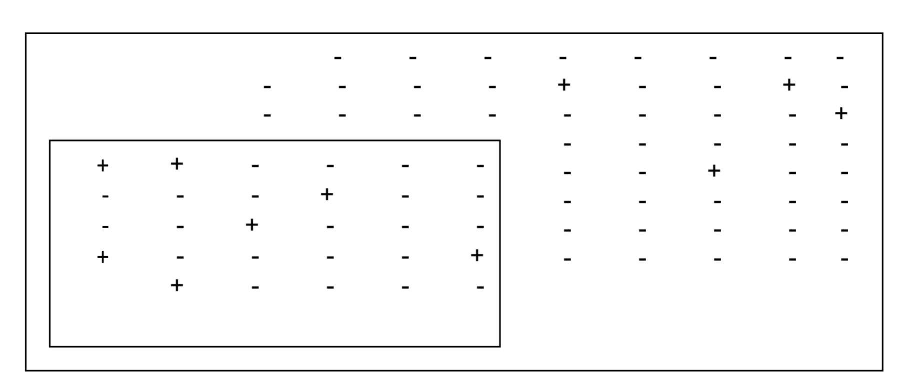
يوجد جانبان أساسيان للمشكلة التي تعالجها نظم تمثيل واسترجاع المعلومات، الجانب الأول هو الجانب الرياضي المتعلق بقدرة النظام على تحقيق أعلى معدلات للاستدعاء والتحقيق في عمليات الاسترجاع، والجانب الثاني هو الجانب الإجرائي المتعلق بقدرة النظام على أداء المهام بفاعلية وتوفير متطلبات سهولة الاستخدام من جانب المستفيدين. وسيتم فيما يلي استعراض كل جانب من هذين الجانبين وتحليله بالتفصيل والتعرف إلى أساليب قياسه:

2.1.1 الجانب الرياضي

يصف الشكل رقم (2.1) مشكلة استرجاع المعلومات، والتي تسعى كل نظم استرجاع المعلومات إلى حلها. ويتضمن الشكل مستطيلين أحدهما كبير والآخر

صغير. يشير المستطيل الكبير في الشكل إلى قاعدة بيانات يتم إعدادها من خلال نظم تمثيل البيانات مثل فهرسة وتكشيف واستخلاص الوثائق التي يتم اختيارها وتحليلها في النظام، بينما يمثل المستطيل الصغير استفسار المستفيد والنتائج المسترجعة. وتمثل علامة (+) في الشكل الوثائق الصالحة التي يرغب المستفيد في استرجاعها من النظام، بينما تمثل علامة (-) الوثائق التي يحكم عليها المستفيد من النظام على أنها غير صالحة. وبالطبع فإن مجموعة الوثائق غير الصالحة (-) لأي استفسار أكبر بكثير من مجموعة الوثائق الصالحة (+) في النظام، بالتالي فإن مشكلة استرجاع المعلومات تتلخص في قدرة النظام على استرجاع أكبر عدد ممكن من الوثائق الصالحة في النظام (+) وأقل عدد من الوثائق غير الصالحة، وبالطبع فإن الحالة المثالية هي استرجاع كل الوثائق الصالحة واستبعاد كل الوثائق غير الصالحة.

وتعتمد الدقة في الاسترجاع بشكل كبير على مدى الدقة في العمليات، والتي تتضمن جزأين رئيسين هما: الجزء الخاص باختيار وتكشيف الوثائق، والجزء الخاص بترجمة احتياجات المستفيدين إلى استراتيجيات بحث تتطابق مع المصطلحات المستخدمة في التعبير عن المحتوى الموضوعي للوثائق. ويمثل المستطيل الأصغر في الشكل رقم (2.1) نتائج البحث في قواعد البيانات. فيوضح المستطيل أنه تم



شكل (2.1) نموذج للجانب الرياضي لمشكلة استرجاع المعلومات

استرجاع 20 وثيقة منها 6 وثائق صالحة (+)، 18 وثيقة غير صالحة (ـ). بالتالي يكون معدل الوثائق الصالحة إلى إجمالي الوثائق المسترجعة 24/ 6 أي 25 ٪. ويستخدم هذا المؤشر لقياس معدل التحقيق Precession Rate الذي يشير إلى مدى الدقة في استرجاع الوثائق الصالحة فقط (Buckland, Fredric , 1994).

عدد الوثائق الصالحة المسترجعة

معدل التحقيق = -------------------------- X 100

عدد الوثائق المسترجعة

ويستخدم معدل الاستدعاء Recall rate للدلالة على استرجاع كل الوثائق الصالحة من قاعدة البيانات، بمعنى آخر معدل الوثائق الصالحة المسترجعة إلى إجمالي الوثائق الصالحة في قاعدة البيانات.

فإذا افترضنا أن قاعدة البيانات تتضمن 100 وثيقة صالحة تم استرجاع 6 منها، يكون معدل الاستدعاء في هذه الحالة (6 / 100) X 100 أي نحو 6٪. ويمكن تحسين معدلات الاستدعاء من خلال توسيع نطاق البحث في النظام باستخدام مصطلحات أكثر شيوعاً أو تردداً في الوثائق، ولكن على الجانب الآخر سوف ينخفض معدل التحقيق عند ارتفاع معدلات الاستدعاء، وذلك لزيادة عدد الوثائق المسترجعة، ما يزيد احتمال ارتفاع عدد الوثائق غير الصالحة.

بالتالي، يتضح أن من أهم عناصر كفاءة نظم استرجاع المعلومات العمل على

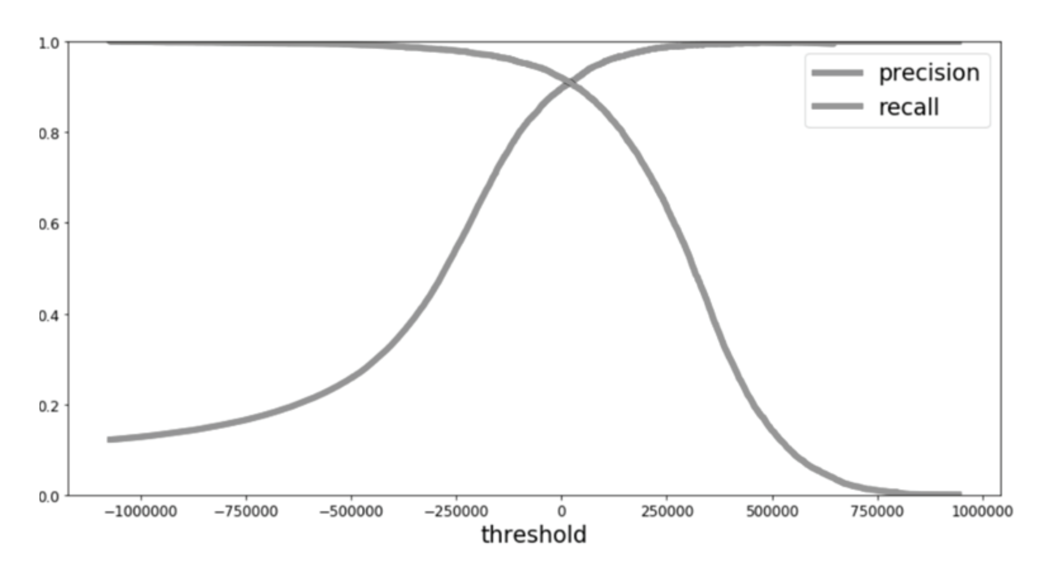
عدد الوثائق الصالحة المسترجعة

معدل الاستدعاء = -------------------------- 100

إجمالي عدد الوثائق الصالحة في النظام

التحسين في معدلات الاستدعاء التي تؤدي بالتبعية إلى انخفاض معدلات التحقيق والعكس صحيح، بمعنى أن ارتفاع معدلات التحقيق يؤدي إلى انخفاض معدلات الاستدعاء. من ثم فإن العلاقة بين الاستدعاء والتحقيق هي علاقة عكسية حتمية كما هو موضح في الشكل (2.2).

ويتضح من الشكل (2.2) أنه توجد علاقة عكسية بين الاستدعاء والتحقيق. وتشير تلك العلاقة إلى أن زيادة معدلات الاستدعاء تعني زيادة عدد الوثائق المسترجعة، وارتفاع احتمالات ظهور وثائق غير صالحة نتيجة لتوسيع نطاق البحث. وعلى الجانب الآخر، فإن تحقيق أعلى معدلات الدقة يتطلب صياغات معقدة لعبارات البحث وتضييق نطاق البحث إلى أقصى درجة ممكنة، ما تقل معه فرص استرجاع عدد كبير من الوثائق، حيث إن الهدف من التحقيق هو الوصول إلى أعلى معدلات الصلاحية في الوثائق المسترجعة.



شكل (2.2) العلاقة العكسية بين الاستدعاء والتحقيق (Buckland, Fredric , 1994)

نموذج افتراضي

إذا افترضنا أن مستفيداً يبحث عن سيارات الدفع الرباعي من فئة تويوتا. وبفحص نظام استرجاع المعلومات تم التوصل لما يلي:

- 50 وثيقة في موضوع السيارات
- 20 وثيقة في موضوع الدفع الرباعي
- 100 وثيقة في الموضوع تويوتا (على افتراض أن المصطلح تويوتا قد يمثل اسم شخص، موديل سيارة، اسم مصنع، أو شركة.. الخ).

وعند الربط بين المصطلحات الثلاثة، قد يسترجع النظام 20 وثيقة صالحة بحد أقصى لهذا الاستفسار. فإذا فحص المستفيد النتائج، ووجد أن هناك 5 وثائق غير صالحة، وعلى افتراض أن النظام يحتوي على 50 وثيقة صالحة.

بالتالي يكون معدل الاستدعاء = (15 / 50) * 100 = 30 ٪

ومعدل التحقيق = (15 / 20) * 100 = 75 ٪

ويلاحظ من هذه النتيجة ارتفاع معدل التحقيق وانخفاض معدل الاستدعاء.

فإذا افترضنا أن المستفيد أراد الحصول على عدد أكبر من الوثائق، فأضاف مصطلح الدفع الكلي إلى مصطلح الدفع الرباعي، وربط بينهما بالمعامل OR لتصبح عبارة البحث كالتالي:

سيارات AND (الدفع الرباعي OR الدفع الكلي) AND تويوتا

وقد أصبح عدد النتائج المسترجعة وفقاً لهذه الاستراتيجية كالتالي:

سيارات = 50 وثيقة

الدفع الرباعي OR الدفع الكلي = 35 وثيقة

تويوتا = 100 وثيقة

ما يعني أنه يوجد 15 وثيقة مكشفة تحت مصطلح الدفع الكلي، وأن خمس وثائق من هذه المجموعة ورد فيها مصطلحا سيارات وتويوتا، بالتالي تكون نتيجة العبارة البحثية كالتالي:

25 وثيقة مسترجعة بحد أقصى عند الربط بين المصطلحات الأربعة وفقاً للعبارة البحثية السابقة. وإذا افترضنا أن عدد الوثائق الصالحة بالنظام كله بعد إضافة المعامل الجديد ارتفع من 50 وثيقة إلى 55 وثيقة. وعند تقييم المستفيد للنتائج المسترجعة (25) وجد أنه توجد 18 وثيقة صالحة و7 وثائق غير صالحة.

بالتالي يكون معدل الاستدعاء والتحقيق هو كالتالي:

الاستدعاء = (18 / 55) * 100 = 32.7 ٪

التحقيق = (18 / 25) * 100 = 72 ٪

ويلاحظ من المعادلة أن معدل الاستدعاء زاد بنسبة 2 ٪ تقريباً، تبعه انخفاض في معدل التحقيق بنسبة 3 ٪ تقريباً، ما يؤكد العلاقة العكسية الحتمية بين الاستدعاء والتحقيق، والتي تأتي كنتيجة منطقية لطبيعة العلاقة، حيث إن ارتفاع الاستدعاء يتطلب توسيع نطاق البحث في حين التحقيق يتطلب تضييق نطاق البحث لتحقيق أعلى معدلات الدقة في النتائج المسترجعة. وتجدر الإشارة إلى أن العلاقة العكسية في الزيادة والنقصان تحدث بشكل نسبي، ولا تسير في اتجاه الزيادة والنقصان المطلق فقط، بمعنى أن الزيادة في الاستدعاء قد تتبعها زيادة في التحقيق ولكن بمعدل أقل في أي منهما.

كما يتضح من الشكل (2.1) أيضاً ظاهرة أخرى من ظواهر نظم تمثيل واسترجاع المعلومات تتمثل في أنه من الممكن توسيع نطاق البحث لاسترجاع كل الوثائق الصالحة (بمعنى تحقيق 100 ٪ استدعاء)، ولكن ذلك سوف يجعل معدل التحقيق منخفضاً جداً، هذا إضافة إلى أنه كلما كبر حجم قاعدة البيانات، انخفض معها معدل التحقيق المحتمل في مثل هذه الحالات. فالمستفيد قد يرغب في فحص مستخلصات 25 وثيقة لكي يصل إلى 5 وثائق صالحة، بينما قد لا يرغب في فحص 100 وثيقة لكي يحصل على 25 وثيقة صالحة، لأن هذا يتطلب جهداً أكبر ووقتاً أطول. بالتالي

فإنه مع زيادة حجم قاعدة البيانات قد يكون من الصعب تحقيق مستوى مقبول من الاستدعاء في مقابل مستوى مقبول من التحقيق. وتوجد العديد من الدراسات التي ركزت على هذه النقطة الجدلية ومازالت هذه النقطة محل خلاف بين الباحثين في مجال استرجاع المعلومات.

ويستخدم لانكستر مصطلح الاستدعاء للدلالة على استرجاع الوثائق الصالحة، أو بشكل أكثر دقة للدلالة على تجنب الوثائق غير الصالحة. كما توجد مقاييس أخرى لقياس أداء البحث في قواعد البيانات. (انظر على سبيل المثال روبرتسون وجونز Robertson & Jones, 1976). بعض هذه المقاييس رياضي بحت، إلا أن الاستدعاء والتحقيق هما أكثر المقاييس استخداماً وتطبيقاً في الأنظمة والدراسات، لما لهما من قدرة على توضيح الصورة العامة لكفاءة نظم تمثيل واسترجاع المعلومات. كما يبدو أنهما مازالا أكثر المقاييس وضوحاً للتعبير عن نتائج البحث، حيث إنهما يقسمان قاعدة البيانات ببساطة إلى قسمين هما وثائق مسترجعة ووثائق غير مسترجعة أو وثائق صالحة ووثائق غير صالحة.

ونظراً للعلاقة العكسية الواضحة بين الاستدعاء والتحقيق تسعى الكثير من الأنظمة إلى استخدام معامل تطبيع البيانات، والذي يعرف بالمعامل F وهو عبارة عن مؤشر لمقياسي الاستدعاء والتحقيق ويتم قياسه وفقاً للمعادلة التالية (Su, 1992):

$$\text{معامل } F = 2\,\frac{\text{التحقيق} \times \text{الاستدعاء}}{\text{التحقيق} + \text{الاستدعاء}}$$

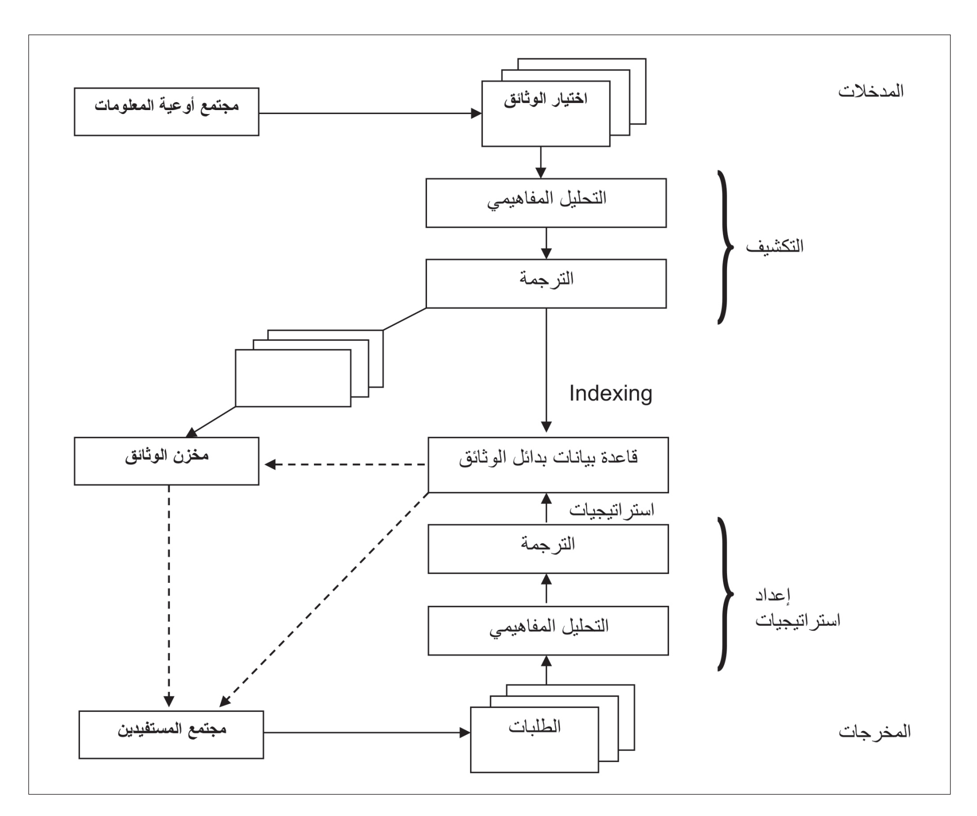
الاستدعاء

◄ 2.1.2 الجانب الإجرائي

تحاول كل نظم تمثيل واسترجاع المعلومات حل المشكلة الإجرائية المتعلقة بآلية عمل نظام تمثيل واسترجاع المعلومات والذي يحاول الإجابة عن السؤال التالي:

كيف يمكن الحصول على المعلومات الصحيحة للمستفيد المناسب في الوقت الملائم، على الرغم من وجود متغيرات أخرى كثيرة (مثل سمات المستفيدين)، تغطية قاعدة البيانات في بيئة نظم تمثيل واسترجاع المعلومات اختلاف أساليب البحث والاسترجاع وخوارزميات معالجة المعلومات.. إلخ.

ويشتمل الشكل (2.3) على نموذج مبسط للمشكلة التي تعالجها نظم استرجاع المعلومات من الناحية الإجرائية:



شكل (2.3) العلاقة العكسية بين الاستدعاء والتحقيق (Buckland, Fredric , 1994)

فالمشكلة الأساسية التي تعالجها معظم نظم استرجاع المعلومات هي مضاهاة احتياجات المستفيدين ببدائل الوثائق المخزنة في قواعد البيانات بنظم استرجاع المعلومات. وتشتمل تلك البدائل على تبسيط للرسائل التي يسعى المؤلفون إلى توصيلها إلى مجتمع المستفيدين والتي تظهر في النصوص أو الوسائط غير النصية التي يقومون بتأليفها في الوقت الذي يتم فيه التعبير عن احتياجات المستفيدين في صورة طلبات يتم توجيهها إلى خدمات المعلومات.

وتقوم نظم استرجاع المعلومات بالتعامل مع إعداد بدائل للنصوص (التي يمكن أن تتراوح بين النص الكامل للوثيقة في شكل إلكتروني أو أجزاء من ذلك النص إلى تسجيلة ببليوجرافية بسيطة تمثل الوثيقة) ويتم تخزينها في قاعدة بيانات يمكن البحث فيها من خلال إحدى أدوات البحث والاسترجاع. ويمكن تخزين قاعدة البيانات في صورة وثائقية أو إلكترونية، ولكنها غالباً ما تتاح عبر شبكة الإنترنت حالياً. أما الأداة التي تستخدم في بحث تلك النظم فيمكن أن تتراوح ما بين النظم التقليدية مثل الفهارس البطاقية أو الكشافات المطبوعة، ولكنها في معظم الأحوال حالياً تتاح من خلال محركات وأدوات البحث المتاحة من خلال شبكة الإنترنت والأجهزة الذكية.

ويتم تجهيز بدائل لطلبات المستفيدين (والتي يتم تمثيلها في شكل مصطلحات يتم الربط بينها من خلال مجموعة من الروابط المنطقية أو تعبيرات نصية أو كيانات)، فعلى سبيل المثال تسمح بعض النظم للباحث بإدخال تفاصيل عن أحد الكيانات المعروفة بأنها صالحة للبحث عن مواد مشابهة لهذا الكيان. ويتم استرجاع بدائل النصوص التي تضاهي بديل الطلب.

ومن أهم المشكلات التي تواجهها مثل هذه النظم أن الرسالة التي يريد المؤلف توصيلها لم يتم التعبير عنها بشكل جيد في النص الذي يعتمد عليه في إعداد بديل الوثيقة، وفي المقابل يمكن أن تكون استراتيجية البحث التي تعبر عن طلب المستفيد قد تم إعدادها بشكل غير جيد ومن ثم لا تضاهي احتياجات المستفيد.

بذلك يمكن القول إن مشكلة استرجاع المعلومات يمكن التعبير عنها بأنها محاولة مضاهاة بدائل احتياجات المستفيدين ببدائل رسائل المؤلفين التي يتم التعبير عنها

في نصوص الوثائق. وترى باتس (Bates, 1996) أن مشكلة استرجاع المعلومات تبدو أكثر تعقيداً مما هي عليه، حيث أشارت إلى أنها مشكلة لا تقتصر على جانب واحد في التعامل مع النظم، فهي تشمل جانبي المدخلات والمخرجات. ولصعوبة التعامل مع جانب المدخلات ركزت معظم الدراسات بشكل أساسي على جانب المخرجات في أنشطة استرجاع المعلومات المتمثل في احتياجات المستفيدين وبدائل الطلبات، واستراتيجيات البحث أكثر من تركيزها على المدخلات المتمثلة في رسائل المؤلفين وبدائل النصوص، وذلك على افتراض أن جانب المدخلات أكثر تعقيداً من جانب المخرجات.

وقد أشار بيلكن (Belkin, 1980) إلى مشكلة استرجاع المعلومات على أنها محاولة مضاهاة بين حالة معرفية مجهولة لصاحب الطلب بحالة معرفية أكثر تماسكاً وتحديداً والمتمثلة في نص المؤلف. ويتمثل دور المكشف في محاولة التنبؤ بأنواع الطلبات التي يمكن أن ترد لطلب وثيقة معينة، والتي تعد في هذه الحالة استجابة جيدة للطلب، ما يحقق رضا المستفيد. ويمكن إنجازها من خلال دور المكشف الذي يحاول تحديد أنواع الوثائق التي تلبي احتياجات مستفيد بعينه في وقت معين.

كما يتضح في الشكل (2.3) أنه يمكن استخدام الخوارزميات في بعض أنشطة استرجاع المعلومات كبديل للتحليل المفاهيمي أو المعالجة البشرية للوثائق. ويتم استخدام ذلك في نظم التكشيف والاستخلاص الآلية وغيرها من العمليات التي تشتمل على معالجات لفئات معينة من الوثائق والمصطلحات مثل بناء استراتيجيات البحث وإعداد شبكات الربط بين المصطلحات، كما هو الحال في المكانز والأنطولوجيات (أدوات معالجة المصطلحات). فكما هو واضح من الشكل يمكن للحاسبات أن تستخدم لمساعدة المكشفين ـ كما هو الحال في معظم قواعد البيانات ومحركات البحث المتاحة عبر الشبكة العنكبوتية، كبديل للمكشفين وذلك في كل أنشطة ومكونات نظم استرجاع المعلومات.

وقبل البدء في مناقشة آليات تمثيل واسترجاع المعلومات بالتفصيل، لابد من التعرض لعملية تمثيل واسترجاع المعلومات للتعرف إليها بدقة.

2.2 عملية تمثيل واسترجاع المعلومات

يقوم أخصائي المعلومات في تلك النظم بتصميم وتنفيذ وصيانة النظم ويقوم المستفيد بإجراء البحث واستقبال النتائج المسترجعة، لذلك فإن أي معلومات يتم استرجاعها من قاعدة البيانات يلعب أخصائي المعلومات دوراً محورياً في تنظيمها وفقاً للغة المستخدمة بالنظام. وكثيراً ما تظهر بعض التناقضات أثناء عملية تمثيل واسترجاع المعلومات، والتي من الممكن أن تؤدي إلى مشكلات كبيرة إذا كانت اللغة المقيدة هي اللغة المستخدمة ويرجع ذلك للأسباب التالية:

أولاً: الاختزال: لأن المعلومات التي يتم تسجيلها في صورة مقالات، دوريات أو تقارير فنية أو أعمال مؤتمرات يتم تمثيلها في صورة ملخصة باستخدام مصطلحات التكشيف Indexing terms وما يشبهها، من ثم فاسترجاع المعلومات الأصلية يبدو من الصعب تحقيقه. فالعملية تشبه هنا تمثيل وثيقة كبيرة بها آلاف الكلمات بعدد محدود من الكلمات، من ثم يكون هذا التمثيل اختزالاً للبعد الخاص بالجحم.

ثانياً: المضاهاة الجزئية: تُعد أي لغة مضبوطة جزءاً من اللغة الطبيعية التي تم كتابة الوثيقة الأصلية بها، لذلك من الصعب أن تحدث مضاهاة كاملة بين كلمة في وثيقة وأخرى مشتقة من مكنز مصطلحات (لغة مضبوطة) لأغراض التمثيل. فمن الممكن أن يكون المكشف قد قام باختيار مصطلح مرتبط أو مصطلح أضيق أو أوسع للدلالة على المفهوم الذي يرغب في التعبير عنه من الوثيقة، وهو ما يجعله غير مطابق كلياً للمصطلح الوارد في الوثيقة.

ثالثاً: عدم الاطراد inconsistency: من التحديات التي يصعب تحقيقها في عمليات التمثيل هو الثبات في تمثيل المعلومات (بما في ذلك عملية تحليل المفاهيم)، والذي يبدو حتمياً وخاصة إذا قام أكثر من شخص أو نظام بأداء المهمة. وقد أشار (كلفردون Cleverdon,1984) إلى أن أكثر المكشفين خبرة يتفقون فقط في حدود 30 ٪ فقط في المصطلحات المستخدمة في التكشيف إذا قاموا بتكشيف نفس الوثيقة، بمعنى أن الاطراد بينهم لا يتجاوز 30 ٪. وفي السياق نفسه وجد (ميتشل

(Mitchell,2003) أن معدلات الاتفاق بين مصطلحات التكشيف باستخدام قائمة رؤوس الموضوعات الطبية MESH [1] في بناء قاعدة بيانات Medline لم يتجاوز نسبة 49 ٪ من المصطلحات المستخدمة في تكشيف الوثائق الطبية. وهو نفس ما توصل إليه محمد (1999) فيما يتعلق بتكشيف الدوريات العربية بقواعد البيانات الوطنية المصرية، حيث توصل إلى أن نسبة الاطراد لا تتجاوز 40 ٪ في مصطلحات التكشيف، على الرغم من التوافق حول الأدوات والسياسات المستخدمة، إلا أن عدم الاطراد يأتي من اختلافات بين المكشفين في عمليات التحليل المفاهيمي والترجمة.

وعلى الجانب الآخر، يحتاج المستفيدون إلى تحويل احتياجاتهم المعلوماتية إلى استفسارات باستخدام لغات نظم تمثيل واسترجاع المعلومات، بحيث يمكن استخدام هذه الاستفسارات في إجراء البحث بقواعد البيانات باستخدام آليات البحث المتاحة. وقد أشار الباحثون منذ القدم إلى مدى تعقد تلك العملية، فقد أوضح بلير ومارون (Blair & Maron, 1985) أنه من الصعب أن يستطيع المستفيد التنبؤ بالكلمات المطابقة تماماً Exact Words أو مزيج الكلمات Word Combination للمصطلحات التي تستخدم تمثيلاً في كل أو معظم الوثائق الصالحة (p.295). وإضافة إلى ذلك، فإن استخدام المصطلحات المضبوطة وإمكانيات البحث (مثل البحث البوليني) سوف يزيد من تلك الصعوبة. وعادة ما يتم استخدام اللغة الطبيعية في البحث بالاعتماد على العبارات والجمل الكاملة التي يتم استخدامها في التواصل في حياتنا اليومية دون أي إجراءات لبناء الاستفسارات (على سبيل المثال لماذا لون السماء أزرق) أصبح أمراً من الممكن البحث عنه على الإنترنت بنفس الطريقة التي يصيغ بها المستفيد استفساره؛ إلا أن الطريق مازال طويلاً أمام الباحثين في هذا المجال، لتوفير آليات لمعالجة اللغة الطبيعية التي تعد أحد أقسام الذكاء الاصطناعي Artificial Intelligence (AI) لإحداث التطوير المنشود في عمليات البحث بالأسئلة المباشرة. وبمعنى آخر، يعتمد نجاح البحث بصفة أساسية على المضاهاة التي تتم بين تمثيل المعلومات بالنظام والاستفسار الذي يتم توجيهه من خلال المستفيد

---

(1) MESH: Medical Subject Headings.

إلى النظام. أي أن عملية البحث تنجح عندما يحدث التطابق بين استفسار المستفيد والمعلومات التي يتم تمثيلها بقاعدة البيانات التي يتم البحث فيها، وفي حالة عدم التطابق لن يستطيع النظام استرجاع النتائج الصالحة.

لذلك، فإن المضاهاة هي الآلية الأساسية بنظم تمثيل واسترجاع المعلومات وكما هو موضح في الشكل (2.3). مع ملاحظة أنه توجد عدة أنشطة بعملية تمثيل واسترجاع المعلومات يمكن أن تؤدي إلى التناقض في المضاهاة. فالهدف النهائي لجودة نظم تمثيل واسترجاع المعلومات هو استخدام كل الطرق والتقنيات الممكنة لتقليل أو حتى القضاء على كل التناقضات التي تظهر أثناء عملية التمثيل والاسترجاع.

2.3 تحديات التمثيل واسترجاع المعلومات

على الرغم من الكمّ الكبير من الدراسات والبحوث التي يتم إجراؤها في مجال نظم تمثيل واسترجاع المعلومات؛ فإنه يوجد مجموعة من التحديات التي فيما يبدوا أنها من الصعب التغلب عليها. فقد قام سوانسون (Swanson,1998) بعرض أفكاره عن التكشيف والاسترجاع الآلي Automatic Indexing & Retrieval قام باستخدام مصطلح سكّه تايلور ويتكار Taylor Whittaker المعروف بمسلمات العجز Postulate of Impotence وحدد 9 مسلمات عجز لا تستطيع نظم تمثيل واسترجاع المعلومات التغلب عليها. على الرغم أن ذلك كان في عام 1988 والذي يشير إلى بدايات عصر الميكنة، إلا أن بعض هذه الصعوبات والتحديات التي وردت في المسلمات التسع لايزال قائماً ونذكر منها على سبيل المثال المسلمات 1، 3، 4، 9. وهذه المسلمات التسع هي:

1. «لا يمكن التعبير عن الحاجة إلى المعلومات بصورة كاملة في صورة طلب بحث؛ حيث لا يمكن صياغة السؤال بصورة دقيقة وبشكل مستقل عن الافتراضات المسبقة التي تكون في ذهن المستفيد، والتي لا حصر لها – كما أنه من المستحيل وصف السياق المعرفي للمستفيد بصورة كاملة، لأنه يشمل، ضمن أمور أخرى، الخلفية المعرفية الخاصة بالمستفيد والطلب».
ويرجع ذلك إلى أن هذه الاحتياجات تنبع أساساً من حالة عدم يقين أو عدم

المعرفة والغموض والالتباس، ومن ثم لا يمكن لتلك الحالة الغامضة أن ينتج عنها سؤال دقيق أو طلب استفسار سليم 100 ٪. وتجدر الإشارة إلى أننا قمنا بدراسة للتغلب على هذا التحدي من خلال ابتكار نموذج تفاعلي لسد الفجوة في حالة عدم اليقين وتحويلها إلى حالة تفاعل تمكن الباحث من الوصول إلى اليقين (انظر محمد، 2013).

2. «لا يمكن توجيه نظام استرجاع معلومات إلى إجراء ترجمة دقيقة لطلب محدد إلى مجموعة مناسبة من مصطلحات البحث. فمصطلحات البحث هي افتراضات واختزالات أو تخمينات لحالة معرفية؛ بالتالي لا توجد قوانين حاكمة لهذا الأمر».

3. «لا يمكن اعتبار الوثيقة صالحة لطلب معلومات بشكل مستقل عن جميع الوثائق الأخرى، التي يجب أن يأخذها المستفيد في الاعتبار. فالصلاحية ليست حكماً ثابتاً، إنما هي عبارة عن أحكام تختلف من سياق لآخر، ومن مستفيد لآخر، ويجب أن تراعي الإطار المعرفي المتغير Shifiting Knowledge Framework».

4. «من المستحيل أن تؤكد أو تنفي أن كل الوثائق الصالحة لاستفسار معين تم الوصول إليها ضمن قائمة النتائج المسترجعة، كما أنه لا يمكن أبداً لأي مستفيد في الممارسة العملية أو من حيث المبدأ فحص جميع الوثائق سواء المسترجعة أو الصالحة بالنظام».

5. «لا يمكن للأجهزة، حتى الآن، أن تتعرف إلى المعنى، بالتالي لا يمكن أن يحدث تطابق كامل بين آليات عمل الأجهزة وما تقوم به من عمليات تكشيف وتصنيف، وأحكام الصلاحية التي يقوم بها البشر. فالنتيجة الطبيعية لذلك: أن بعض المكشفين طوال الوقت، وجميع المكشفين في بعض الأوقات، لا يمكنهم تحقيق التطابق مع ما يمكن للمستفيدين إضافته إلى عمليات التكشيف والتصنيف أثناء إجراء أحكام الصلاحية». وهو ما دفع الباحثين إلى ابتكار أساليب التوسيم الاجتماعي Social Tagging.

6. «معدل تردد المصطلحات Word-occurrence لا يمكن أن يمثل المعنى أو حتى يكون بديلاً له، ومع ذلك فإن هذه البيانات يمكن أن تستخدم لتحقيق نجاح عرضي في عملية البحث، في الإشارة إلى أو لتحديد المناطق المهمة في النص التي يمكن للمستفيد أن يستخدمها في البحث عن المعنى أو الحكم على الصلاحية».

7. «لا يمكن تقييم قدرة نظام استرجاع المعلومات على دعم عملية تكرارية من خلال أحكام الصلاحية المفردة التي يجريها المستفيد مرة واحدة لعمليات متكررة single-iteration human relevance judgment، فالعمليات المتكررة تحتاج إلى معايير جديدة للحكم مثل قدرة النظام على تحفيز المراجعة الإبداعية للسؤال أو الاستفسار أثناء تفاعل المستفيد مع النظام».

8. «لا يمكن للنظام أن يجمع بين أحكام الصلاحية البشرية والآلية، فالنظام إما أن يستخدم أحكام صلاحية بشرية دقيقة أو إجراءات ميكانيكية فعالة للغاية، لكن ليس كليهما معاً».

9. «باختصار تشير المسلمات الثماني الأولى إلى أن تحقيق الفعالية والكفاءة الكاملة باطراد من خلال إجراءات التكشيف والاسترجاع الآلي أمر غير ممكن من الناحية العملية».

فالمشكلة المفاهيمية Conceptual Problem لاسترجاع المعلومات كما وصفها ساونسون (Swanson, 1998) هي من أكثر الأمور أهمية في فهم وتطوير مجال استرجاع المعلومات. فالفحص الدقيق لعملية تمثيل واسترجاع المعلومات يوضح أن هذا المجال يتضمن، كما أوضحنا مسبقاً، مضاهاة للمصطلحات وليس بحثاً عن المفاهيم في البيئة الرقمية. فعندما يكون المصطلح البحثي المواصلات العامة Public Transpotation على سبيل المثال لا يمكن استرجاع الوثائق التي تتناول موضوعات الطرق، الأتوبيسات ومترو الأنفاق؛ إلا إذا كان هناك علاقات تربط بين تلك المصطلحات في قاعدة بيانات من خلال قائمة المصطلحات المضبوطة أو أدوات الربط الدلالي. من ثم فالمشكلة المفاهيمية لاسترجاع المعلومات والتي يطلق

عليها مشكلات المعنى problems of meaning لا تقل عمقاً في جوهرها عن غيرها من أشكال السلوك الذكي Intelligent behavior (p.96)، وهو الموضوع الذي ركزت عليه دراسة بناء المفاهيم وإشكالية دلائل المصطلحات التي قام بها مؤلف هذا الكتاب لوضع آلية لتفاعل المستفيدين مع النظام تُمكن من التغلب على المشكلة المفاهيمية عند بناء الاستفسارات (محمد، 2013).

من ثم يمكن القول بإيجاز، إن مشكلة استرجاع المعلومات التي تم اختزالها في عملية مضاهاة مصطلحات ومدى مطابقة الكلمات البحثية للمصطلحات الكشفية، هي أعمق بكثير من مجرد عملية مضاهاة سطحية إلى مضاهاة في الدلالات والمعاني والسياقات.

ونختتم هذا الفصل بسؤال مهم: هل يمكن أن ينتقل استرجاع المعلومات في يوم ما من مجرد أداة لمضاهاة المصطلحات إلى ابتكار آليات للبحث عن المفاهيم؟ الإجابة عن هذا السؤال تم اختبارها ومحاولة الرد عليها بقوة من خلال التجربة والخطأ (Swanson,1998). ويمكن الوصول إلى إجابة كاملة عنها في كتاب الويب الدلالي (محمد وآخرون، 2018).

المصادر

- محمد، خالد عبد الفتاح؛ عتمان، إسماعيل؛ النشرتي، مؤمن؛ حسنين، رجب (2018) الويب الدلالي: المفاهيم والتطبيقات، الرياض، دار المتنبي، 312 ص.
- محمد، خالد عبد الفتاح (أكتوبر 2013). بناء المفاهيم وإشكالية دلائل المصطلحات في تفاعل المستفيدين مع نظم استرجاع المعلومات. مجلة المكتبات والمعلومات العربية. س (33). ع (4): 35 ـ 8
- محمد، خالد عبد الفتاح (1999). التعاون والتنسيق بين مراكز المعلومات القطاعية والشبكة القومية للمعلومات في مصر، جامعة القاهرة (أطروحة ماجستير)، 300 ص.

- Bates, Marcia J. “An exploratory paradigm for online information retrieval.” Intelligent Information Systems for the Information Society. Amsterdam: North-Holland (1996): 91-99.

- Belkin, Nicholas J. "Anomalous states of knowledge as a basis for information retrieval." Canadian journal of information science 5.1 (1980): 133-143.
- Boley, D., Gini, M., Hastings, K., Mobasher, B., & Moore, J. (1998). A client-side Web agent for document categorization. Internet Research, 8(5), 387-399.
- Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28(3), 289-299.
- Buckland, Michael, and Fredric Gey. "The relationship between recall and precision."Journal of the American society for information science45.1 (1994): 12-19.
- Chu, H. (1997). Hyperlinks: How Well Do They Represent the Intellectual Content of Digital Collections?. In Proceedings of the ASIS Annual Meeting (Vol. 34, pp. 361-69).
- Cleverdon, C. (1984). Optimizing Convenient Online Access to Bibliographic Databases. Information services and Use, 4, 37-47.
- Craven, T. C. (1986). String indexing. Academic Press.
- Fugmann, R. (1993). Subject analysis and indexing: theoretical foundation and practical advice (Vol. 1). Indeks Verlag Dr. Ingetraut Dahlberg.
- Garfield, E. (1965, December). Can citation indexing be automated. In Statistical association methods for mechanized documentation, symposium proceedings (Vol. 269, pp. 189-192). Washington, DC:
- Jones, K. S. (2007). Automatic summarizing: The state of the art. Information Processing & Management, 43(6), 1449-1481.
- National Bureau of Standards, Miscellaneous Publication 269.
- Kelly, Brian. (2005).RSS: Moe than just news feeds. New review of information Network,11(2), 219-227.
- Kipp, M. E. (2007). @ toread and cool: Tagging for time, task and emotion.
- Lerner, R. M. (2004). At the forge: syndication with rss. Linux Journal, 126(8).
- Lesk, Michael. (1997). Practical digital libraries: Books, Bytes and bucks. San Francisco: Morgan Kaufmann.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.
- Malin, M. V. (1968). Science Citation Index: a New concept in indexing. Library Trends, 16(3), 374-374.
- McKiernan, G. (2001). Beyond bookmarks: schemes for organizing the web.
- Meadow, C. T., Boyce, B. R., & Kraft, D. H. (1992). Text information retrieval systems (Vol. 20). San Diego, CA: Academic Press.

- Mitchell, Joyce A., et al. "Gene indexing: characterization and analysis of NLM's GeneRIFs." AMIA Annual Symposium proceedings. Vol. 2003. American Medical Informatics Association, 2003.
- o'reilly, tim. (2005). / What is Web 2.retrieved October 4, 2009,from oreilly. Com/ web2/ archive/ what – is –web-20.html.
- Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. Journal of the American Society for Information science, 27(3), 129-146.
- Rowley, J. (1992). Organizing knowledge: an introduction to information retrieval. Gower.
- Smith, G. (2008). Information architecture: Tagging: Emerging trends. Bulletin of the American Society for Information Science and Technology, 34(6), 14-17.
- Spark johns, Karen, and Endres- Niggemeyar, Brigitte (EDS.). (1995). Summarizing text [Special issue]. Information processing & management, 31(5).
- Su, Louise T. "Evaluation measures for interactive information retrieval." Information Processing & Management 28.4 (1992): 503-516.
- Swanson, Don R. "Historical note: Information retrieval and the future of an illusion." Journal of the American Society for Information Science 39.2 (1998): 92-98.
- Vander Wal, T. (2007). Folksonomy.retrived November 20.2008.from vanderwal.net/ folksonomy.html
- Xu, C., & Chu, H. (2008). Social tagging in China and the USA: A comparative study. Proceedings of the American Society for Information Science and Technology, 45(1), 1-9.
- Yang,Y.(1999).Anevaluationofstatisticalapproachestotextcategorization. Information retrieval, 1(1-2),69-90.

الفصل الثالث

تمثيل المعرفة:

قضايا أساسية

مقدمة

تتنوع أشكال الوثائق وأنواع مصادر المعرفة التي تعد الناقل الأساسي للمعلومات، حيث تحمل المعلومات التي يتم إنتاجها لأغراض تداول المعرفة منها أعمال المؤتمرات، مقالات الدوريات، التقارير الفنية.. إلخ. وتحتاج هذه الوثائق إلى أن يتم تمثيلها قبل إتاحتها للبحث والاسترجاع، فلا يمكن استرجاع المعلومات التي تتضمنها الوثائق بالاعتماد عليها فقط؛ حيث يحتاج نشاط استرجاع الوثائق إلى بدائل لتلك الوثائق والتي عادة ما تأخذ أشكالاً متنوعة مثل الكشافات، المستخلصات، والملخصات، وغيرها. ويتم التعبير عن تمثيل الوثائق في هذا الكتاب للإشارة إلى جوهر الوثيقة أو المحتوى الموضوعي باستخدام آلية معينة بمصطلح تمثيل المعرفة، على الرغم من أن عملية التمثيل تركز على مخرجات المعرفة التي يتم نشرها في صورة وثائق وأوعية معلومات متنوعة. وقد تم استخدام مصطلح تمثيل المعرفة في هذا الكتاب للدلالة على تمثيل الوثائق التي تعد مخرجات المعرفة الحقيقية والتي تشكل الذاكرة الخارجية للإنسان في مقابل الذاكرة الداخلية، كما تشير إلى كل العمليات الفنية التي تتم على أوعية المعلومات ومنها التكشيف (الهجرسي، 1991).

وتجدر الإشارة إلى أن المنتج النهائي من الممكن أن يأخذ أشكالاً متنوعة، فمن الناحية النموذجية يجب أن تتم عملية تمثيل الوثائق بسهولة وفعالية من خلال إجراءات التمثيل التي سنتناولها بالتفصيل. وقد أشار ليسك (Lesk, 1997, P99 - 100) إلى ما يلي: إذا كان من الممكن تمثيل المعرفة بطريقة واحدة يمكن من خلالها تنظيم الأفكار في مواضع ثابتة، وإذا كان المستفيد على دراية بتلك الطريقة ويمكنه توجيه الاستفسارات بطريقة تتماشى مع تلك الآلية؛ فإن عملية الاسترجاع الموضوعي سوف تعمل بثبات

واطراد، لكن من الناحية العملية من المستحيل أن يتم استخدام طريقة واحدة لتمثيل المعرفة تخدم كافة الأغراض؛ علاوة على ذلك فإن تطبيق عملية التمثيل باطراد ودقة مازال يواجه العديد من التحديات من وجهة نظر أخصائي المعلومات، حتى لو كان اختيار طريقة التمثيل لا يمثل تحدياً، فإن بعض طرق التمثيل مثل المستخلصات لا تستخدم طريقة واحدة ثابتة في التمثيل. لذلك فإن تمثيل مخرجات المعرفة في جوهره يحمل كثيراً من التحديات والتعقيدات، وسنتناول فيما يلي الآليات المختلفة المستخدمة في تمثيل مخرجات المعرفة في صورة بدائل لتلك المخرجات.

3 طرق التمثيل

توجد أساليب متنوعة لتمثيل المعلومات تشمل كل الآليات التي تستخدم في بناء مؤشرات أو بدائل للوثائق. ويستعرض الجزء التالي الأساليب المتنوعة للتمثيل والتي تشمل التكشيف، التصنيف أو التقسيم إلى فئات، التوسيم الاجتماعي، التلخيص، الملخص الوافي للموقع.

3.1 التكشيف Indexing

يُعد التكشيف أحد أنماط تمثيل مخرجات المعرفة التي تم استخدامها على نطاق واسع من جانب الأخصائيين عبر العصور، ويعتمد التكشيف على استخدام مصطلحات (مثل الكلمات والعبارات) سواء كانت بالاشتقاق أو بالتعيين للتعبير عن الأوجه المهمة للوثيقة الأصلية.

وعادة ما يُنظر إليه على أنه العملية التي يتم من خلالها إعداد كشاف يساعد على الوصول إلى التفاصيل الدقيقة للوثائق. وبتجريد المصطلحين تكشيف وكشاف نجد أنهما مشتقان من أصل لغوي واحد وهو «كشف» وتشير القواميس اللغوية إلى أن (كَشَفَ الشيء) يعني أزال الغطاء عنه أو رفع عنه ما يواريه. وقد دخلت كلمة الكشاف اللغة الإنجليزية في العصور الوسطى وتتكون من مقطعين هما In – dex وتشير In إلى ما بداخل الشيء أما Dex فتعني «يشير إلى» أو «يلفت الانتباه إلى» أو

«يدل على وجود شيء». وقد استخدمت كلمة تكشيف في اللغة الإنجليزية بمعنى إعداد كشاف أو إدخال كلمة في كشاف، ثم لحقتها كلمة مُكَشِّف Indexer وتشير إلى الشخص الذي يقوم بإعداد الكشاف.

ويتضح من ذلك أن المعنى اللغوي لكلمة كشاف سواء في اللغة العربية أو في اللغة الإنجليزية يشير إلى إظهار الشيء أو كشف النقاب عنه مع ملاحظة أن اللغة الإنجليزية أظهرت معاني أخرى للكلمة منها قائمة تسبق الكتاب، وقائمة في نهاية الكتاب تضم الأسماء والموضوعات كما تشير إلى أماكن ورودها في النص. **(حسام الدين، 1994)**

أما المعنى الاصطلاحي لكلمة كشاف فنجد له تعريفات متعددة منها تعريف (عبدالهادي، 2005) الذي عرف الكشاف على أنه دليل محتوى المواد التي يحللها أو يكشفها بوساطة دوال معينة ويحدد موضعها أو موقعها بوساطة روابط معينة. كما يعرفه على أنه عبارة عن قائمة أو دليل بمحتويات المواد التي يكشفها بهدف تحديد المفاهيم التي تعالجها والموضوعات التي تعبر عن هذه المفاهيم والأماكن التي وردت فيها في النص.

التكشيف هو تلك العملية الفنية التي ينتج عنها إعداد الكشافات. ويشير لانكستر إلى أن **التكشيف** هو عملية تحليل المفاهيم Conceptual Analysis المرتبطة بمصادر المعلومات التي يتم تكشيفها ونقل هذه المفاهيم إلى مصطلحات تعبر عن المحتوى الموضوعي للوثيقة Document Aboutness من خلال الاعتماد على لغات التكشيف.

ويتراوح عدد المصطلحات الكشفية التي تستخدم للدلالة على وثيقة معينة ما بين عدد محدود من الكلمات بقاعدة بيانات ببليوجرافية إلى مئات الكلمات بنظم النصوص الكاملة. وتنقسم عملية التكشيف التي يتم فيها التعبير عن المحتوى الفكري للوثيقة إلى مرحلتين أساسيتين هما:

- التحليل المفاهيمي Conceptual Analysis
- والترجمة Translation.

وبصورة أكثر تحديداً، يتم في إطار عملية التحليل المفاهيمي تحديد المفاهيم الأساسية التي تتناولها الوثيقة، بينما يتم في مرحلة الترجمة تحويل المفاهيم التي تم تحديدها إلى مصطلحات كشفية بالاعتماد على لغة تكشيف محددة مسبقاً.

ويعرف (عبدالهادي، 2005) عملية التكشيف بأنها عملية خلق أو إيجاد المداخل في الكشاف أو إعداد المداخل التي تساعد على الوصول إلى المعلومات في مصادرها وهي تتضمن 4 عمليات فرعية هي:

1. الفحص الدقيق لأوعية المعلومات للتعرف إلى ما تشتمل عليه من أفكار ومعلومات.
2. تحليل المحتوى الموضوعي للوثائق للتعرف إلى المفاهيم التي تتناولها.
3. تحويل أو نقل المفاهيم إلى مصطلحات مشتقة من لغة التكشيف أو من الوثائق ذاتها.
4. إضافة الروابط التي تعبر عن مكان وجود كل وحدة من الوحدات التي تم تكشيفها داخل المجموعة.

وقد استخدم بعض الباحثين مصطلحات أخرى للدلالة على عملية التكشيف ومصطلحات التكشيف دون تمييز واضح بينها. على سبيل المثال مصطلحات مثل مؤشرات المحتوى Indicators Of Content للدلالة على المصطلحات، بينما يُنظر إلى عملية التكشيف على أنها عملية تحديد المحتوى والمؤشرات الدالة عليه والعلاقات التي تربط بين المؤشرات في الوثائق، بينما يفضل كونر Connor استخدام مصطلح مثل إعداد إشارات Pointing ويشير إلى مصطلحات التكشيف على أنها Pointers وينظر لعملية التكشيف على أنها عملية تحديد إشارات تصف مضمون الوثائق (Lancaster et el., 1991). وتعد الكشافات المخرج الأساسي لعملية تمثيل المعلومات عن طريق التكشيف سواء تمت تلك العملية بطريقة آلية أو يدوية.

3.1.1 أهمية الكشافات

الكشافات أو قواعد البيانات الببليوجرافية هي إحدى الأدوات الأساسية لاسترجاع المعلومات. وأدوات الاسترجاع بصفة عامة هي عبارة عن نظم تم إعدادها لتيسير سبل إتاحة المعلومات. وتتضمن هذه الأدوات تسجيلات ببليوجرافية تعد بدائل للوثائق أو أوعية المعلومات. وتعمل أدوات الاسترجاع على تنظيم أكبر قدر ممكن من أوعية المعلومات التي يتم إنتاجها في جميع أنحاء العالم. ففي سنة 1892 كان كل من بول أتليت تلين Paul-otelt وهنري لافونتين Henry Lafonteen يحلمان بتنظيم مؤتمر دولي بهدف التخطيط لإنشاء نظام دولي للضبط الببليوجرافي Universal Bibliographic Control (UBC) . وكانت معظم الجهود في تلك الفترة تتجه نحو بناء كشافات بالإنتاج الفكري في العلوم والتكنولوجيا.

ويمكننا تخيل أهمية الكشافات أو قواعد البيانات الببليوجرافية إذا تصورنا مقدار الجهد والوقت والكُلفة التي يحتاج إليها الباحث الذي يريد الوصول إلى معلومة وردت في مقالة معينة أو يريد تجميع الإنتاج الفكري حول نقطة معينة يريد إجراء بحث حولها أو باحث يريد الوصول إلى خبر ورد في صحيفة.. أو غيره. بالطبع فإن هذه العملية دون وجود أدوات تيسر هذه العملية سوف تكون مستحيلة في كثير من الأحيان.

بالتالي فإن أهمية الكشافات تأتي مما توفره من سبل وصول إلى المكونات والجزيئات الدقيقة لأوعية المعلومات من كتب ودوريات وغيرها بدرجة عالية من الدقة والسهولة وفي أقل وقت ممكن. ويمكن تلخيص وظائف الكشافات وقواعد البيانات بصفة عامة فيما يلي:

1. حصر الإنتاج الفكري حول موضوع معين وتنقيته لاختيار المواد المهمة بالنسبة للمستفيدين.

2. توفير مداخل وصول منهجية متعددة ومتنوعة لكل وحدة من وحدات المعلومات التي يتم تكشيفها.

3. توفير سبل وصول إضافية من خلال المداخل الإضافية والإحالات وطرق البحث المتنوعة التي توفرها هذه الأدوات.

4. تجميع المصادر المتشابهة معاً في مكان واحد رغم وجودها مبعثرة في الإنتاج الفكري، ما يساعد على الكشف عن العلاقات بين الموضوعات والمفاهيم والمصادر والمؤلفين والدوريات.

5. تساعد الكشافات الموضوعية على التعرف إلى تطورات البحث في مجال موضوعي معين والعلاقات الجديدة بين الموضوعات الحديثة والقديمة.

6. تساعد الكشافات على التعرف إلى المصطلحات المستخدمة في المجالات الموضوعية والتمييز بين المصطلحات المستخدمة وغير المستخدمة والعلاقات بين هذه المصطلحات، وتستمد الكشافات هذه الميزة من أدوات التكشيف وخاصة المكانز.

3.1.2 نظام التكشيف
Indexing System

تتم عملية التكشيف وفقاً لنظام محدد يعرف بنظام التكشيف Indexing System وهو عبارة عن مجموعة من الوحدات التي تتكامل مع بعضها بعضاً بغرض إنتاج الكشافات أو قواعد البيانات. تشمل هذه الوحدات مجموعة القواعد والإجراءات «اليدوية أو الآلية» التي تضبط وتحكم عملية التكشيف، هذا إضافة إلى مجموعة التجهيزات والأدوات اللازمة للتكشيف، والجانب البشري في عملية التكشيف المتمثل في مجموعة المكشفين.

ويمكن القول إن نظام التكشيف يشتمل على المكونات الثلاثة لأي نظام معلومات وهي كالتالي:

3.1.2.1 المدخلات

وتعد المجموعات والمقتنيات التي تمثل المحتوى الفكري الذي يسعى نظام التكشيف إلى تيسير آليات للوصول إليه أهم مدخلات أي نظام للتكشيف، كما تشتمل المدخلات أيضاً على المكشفين والتجهيزات اللازمة لعملية التكشيف.

- **المجموعات**

تشتمل على مجموعة الوثائق التي يتم تكشيفها، ولابد أن تخضع عملية اختيار هذه المجموعات لعمليات فحص دقيقة، حيث إن نظم التكشيف عادة ما تتعامل مع أنواع معينة من الوثائق يطلق عليها الوحدات الصغيرة لأوعية المعلومات أو الميكروميديا Micromedia والتي تشمل أوعية معلومات مثل مقالات الدوريات، فصول الكتب، أعمال المؤتمرات، التقارير الفنية، براءات الاختراع.. الخ. وعادة ما تعمل معظم نظم التكشيف في إطار محدد ودقيق، حيث يتم تجميع أوعية المعلومات التي تدخل في نطاق هذا الإطار سواء كان إطاراً موضوعياً أو شكلياً أو جغرافياً. ويوجد ثلاثة أنماط من أنظمة التكشيف من حيث تغطية المجموعات هي كالتالي:

1. نظم التكشيف التي تغطي نطاقات جغرافية (عالمية، أقليمية، محلية).
2. نظم التكشيف التي تغطي قطاعات معرفية محددة ومجالات موضوعية متخصصة .
3. نظم التكشيف التي تغطي أشكالاً محددة من الوثائق مثل الرسائل الجامعية، براءات الاختراع، الخرائط والوسائط المتعددة.. إلخ.

ومن الجدير بالذكر أن نظم التكشيف العالمية تعتمد في الأصل على الجهود المحلية الرامية إلى تجميع الإنتاج الفكري الوطني، حيث إن تجميع الإنتاج الفكري العالمي كان ومازال أحد الأفكار الرئيسة لمؤسسات المعلومات الدولية مثل الاتحاد الدولي للمكتبات والمعلومات International Federation for Library Association and Institutions - IFLA إلا أنها وجدت أن تحقيق هذا الهدف أمر غير ممكن

وغير عملي في الوقت نفسه، دون التعاون من جانب الحكومات المحلية. لذلك سعت الأمم المتحدة من خلال اليونسيست إلى إنشاء شبكات معلومات محلية في الدول النامية حتى يمكنها المشاركة في حصر وتجميع الإنتاج الفكري الوطني في المجالات العلمية المختلفة إلى جانب المشاركة في البرامج الدولية للمعلومات.
ولعل أبرز نماذج نظم التكشيف العالمية حالياً تتمثل في أدوات البحث التالية:

ISI WEB OF SCIENCE

SCOPUS

GOOGLE SCHOLAR

وتتنافس هذه الأنظمة الثلاثة على تكشيف أكبر قدر من الإنتاج الفكري العالمي وتوفير أدوات لقياس جودة وكفاءة المخرجات العلمية للمؤسسات والدول والجامعات والأفراد والمصادر (الدوريات والمؤتمرات.. إلخ).

- **التجهيزات**

تشمل التجهيزات كل ما يدخل في عملية التكشيف من أجهزة وأدوات ومعايير وقواعد وإرشادات واستمارات وغيرها من التجهيزات اللازمة لعملية التكشيف. وتشمل الأجهزة الداخلة في نظم التكشيف اليوم، حاسبات آلية بأنواعها المختلفة وبرامج متخصصة في عمليات بناء الكشافات واسترجاع المعلومات. وتجدر الإشارة هنا إلى أن هناك نظم تكشيف آلية يمكنها أن تؤدي عملية التكشيف الكامل للوثائق دون الحاجة إلى مكشفين أو لغات تكشيف، حيث إن هذه النظم عادة ما تعتمد على استخدام اللغة الطبيعية للوثائق. أما الأدوات الداخلة في عملية التكشيف فتشمل لغات التكشيف، قواعد الفهرسة، خطط التصنيف، القواميس والمعاجم، سياسات التكشيف.. إلخ.

وتعد القواعد والمعايير من أهم العناصر التي تضبط عملية التكشيف، فهناك مجموعة من المواصفات القياسية التي يتم تطبيقها في نظم التكشيف، ومن أمثلة هذه المواصفات: المواصفة الأمريكية التي صدرت عن الجمعية الأمريكية لعلم المعلومات American Society for Information Science:ASIS ورقمها Z39.41968 والمواصفة

التي صدرت عن المعهد البريطاني للمعايير British Standards Institution بالمملكة المتحدة، والتي تحمل رقم 1976ـB93700 وتحدد هذه المواصفات القياسية مفهوم التكشيف ومخرجات عملية التكشيف والإجراءات المتبعة في عمليات التكشيف ومكونات نظم التكشيف.

- **المكشفون Indexers**

المكشف هو الشخص الذي يقوم بعملية التكشيف، ولابد أن تتوافر في هذا المكشف مجموعة من المؤهلات والخبرات والمهارات التي تمكنه من القيام بعملية التكشيف على أكمل وجه. ولعل أهم المؤهلات التي يجب توافرها في المكشف هو التخصص الموضوعي أو الإلمام الدقيق بالمصطلحات والبناء المعرفي للمجال الموضوعي للوثائق التي يقوم بتكشيفها، بمعنى أن يكون المكشف قادراً على التعامل مع المجال الموضوعي لنظام التكشيف.

ويرى ماثيس (Mathes, 1998) أن عمليات التكشيف التي يتم فيها تحديد واصفات البيانات يمكن أن تقوم بها إحدى الفئات التالية:

- **المكشفون Indexers**

وغالباً ما تعتمد هذه الفئة على اللغات المضبوطة في اختيار وانتقاء المصطلحات الكشفية، وعلى الرغم من تميزها بالجودة العالية والدقة في عمليات تحديد المصطلحات، إلا أن هذه العملية عادة ما تكون مكلفة وتستغرق وقتاً وجهداً كبيرين؛ الأمر الذي يجعل من الصعب الاعتماد عليها بصورة كاملة مع النمو الهائل في المحتوى الذي حدث مع انتشار تطبيقات الإنترنت.

1. **المؤلفون Authors**: المؤلف هو المنشئ الأصلي للوثائق المراد وصفها وتكشيفها. ولكن واصفات بيانات المفهرسين والمؤلفين تشترك في مشكلة أساسية وهي أن المستفيد النهائي من الوثيقة غير متصل بعملية التكشيف هذه أو منعزل عنها تماماً. ولذلك ظهر الاتجاه الثالث، ألا وهو التكشيف من خلال المستفيدين.

2. **المستفيدون Users**: ظهر هذا النوع من التكشيف وانتشر في أواخر عام 1990م من خلال مدونات الويب Web Blogs؛ حيث توفر هذه المدونات روابط Links يتم عرضها جنباً إلى جنب مع تعليقات المستخدمين (أي مقترنة بها)، ويعتمد هذا النوع من التكشيف على اللغة الطبيعية.

يرى البعض أنه من الصعب أن يقوم مكشف غير متخصص بتكشيف وثائق متخصصة في الفيزياء النووية، وفي الوقت الذي لا يعرف فيه هذا الشخص أي شيء عن علم الفيزياء وعلاقة هذا المجال الموضوعي بالمجالات الأخرى. كما يرى البعض أيضاً أنه من الصعب أن يقوم شخص بالتكشيف دون دراسة علمية لإجراءات وآليات التكشيف.

وتوجد وجهتا نظر في هذه الناحية: الأولى ترى ضرورة أن يعمل المتخصصون الموضوعيون على تكشيف أوعية المعلومات في مجالاتهم الموضوعية المتخصصة بعد تدريبهم على أساليب ومبادئ التكشيف. وهذا هو النموذج الأكثر تطبيقاً في معظم أنظمة التكشيف المتخصصة، وقد أوضح محمد (1999) أن 80٪ من المكشوف في مراكز المعلومات القطاعية التي تتولى بناء قواعد البيانات الببليوجرافية المصرية من المتخصصين موضوعياً الذين تم تدريبهم على أساليب التكشيف.

أما الاتجاه الثاني فيرى أنه من الممكن لأخصائيي المعلومات خريجي أقسام المكتبات والمعلومات، أن يقوموا بعمليات التكشيف إذا ما أحسنوا الاستفادة من الأدوات المتاحة لديهم من قواميس متخصصة ولغات تكشيف وخطط تصنيف وغيرها من الأدوات التي تمكنهم من التعرف إلى علاقة الموضوعات ببعضها بعضاً، والمصطلحات المتخصصة في المجالات الموضوعية التي يعملون على تكشيفها. والحقيقة أن لكل وجهة نظر ومزاياها وعيوبها، وإن كان من الأفضل المزج بين الاتجاهين في عمليات التكشيف بغرض الاستفادة من الخبرات الموضوعية إلى جانب الخبرات المهنية، حيث إن عملية التكشيف ليست مجرد مجال علمي يمكن ممارسته بسهولة وإنما هي مهنة بها الكثير من الجوانب العلمية إلى جانب العمليات الفنية التي تحتاج إلى مهارات خاصة تتعلق باستخدام أدوات ومعايير الفهرسة

والتصنيف والتكشيف، إضافة إلى دراسة احتياجات المستفيدين من النظام سواء الحالية أو المتوقعة، كما تتطلب قدراً كبيراً من الثقافة والفهم للعلاقات المتشابكة بين مجالات المعرفة البشرية.

وإلى جانب المؤهلات التي ينبغي أن تتوافر في المكشف لابد أن يتمتع المكشف بمجموعة من المهارات تشمل القدرات اللغوية وإمكانيات التعامل مع الحاسب الآلي وشبكات المعلومات التي تمكنه من نقل وتبادل التسجيلات الببليوجرافية مع النظم الأخرى، وإدارة النظام والتعامل مع قضايا المستفيدين المتعلقة بالدعم الفني وتدريب المستفيدين والرد على الاستفسارات.

3.1.2.2 عمليات التحليل والتكشيف

الجانب الثاني من جوانب نظام التكشيف يتمثل في مجموعة الإجراءات التي تتم من خلالها عملية التكشيف نفسها وتشتمل على خطوتين أساسيتين هما:ـ

- التحليل المفاهيمي
- الترجمة

وستتم مناقشة إجراءات التحليل والتكشيف بالتفصيل لاحقاً.

3.1.2.3 المخرجات

تعد الكشافات وقواعد البيانات ونشرات الاستخلاص أهم مخرجات أي نظام تكشيف واسترجاع معلومات، هذا إلى جانب ما تتضمنه هذه النظم من معالجة لاستفسارات المستفيدين من أجل إجراء البحث عن الوثائق المناسبة لهذه الاستفسارات.

ويشتمل نظام التكشيف على العديد من النظم الفرعية الداخلة في تكوينه، والتي تتفاعل معاً في منظومة واحدة من أجل تلبية احتياجات المستفيدين. ويتيح نظام

التكشيف طرقاً متنوعة لخزن واسترجاع المعلومات التي يمكن من خلالها تلبية احتياجات المستفيدين من النظام بغرض تيسير سبل بحث واسترجاع المعلومات.

3.1.3 التكشيف ونظم تمثيل واسترجاع المعلومات

أشار لانكستر إلى أن نظام استرجاع العلومات يتكون من 6 نظم فرعية هي: (لانكستر، 1997)

1. النظام الفرعي لاختيار الوثائق
2. النظام الفرعي للتكشيف والتحليل
3. النظام الفرعي للغة التكشيف
4. النظام الفرعي للبحث
5. النظام الفرعي الخاص بالتفاعل بين المستفيد والنظام
6. النظام الفرعي الخاص بالمضاهاة

يقع النظام الفرعي للتكشيف في محطتين أساسيتين من محطات العمل في نظم تمثيل واسترجاع المعلومات هما:

- النظام الفرعي للتكشيف والتحليل.
- النظام الفرعي للغة التكشيف.

بالتالي يتضح أن التكشيف يشكل محور نظام تمثيل واسترجاع المعلومات، لأن هذا النظام يعتمد بشكل كبير على المضاهاة بين ناتج عملية التكشيف المتمثل في المصطلحات التي تعبر عن احتياجات المستفيدين، وعملية تحليل الاستفسارات المتمثلة في استراتيجية البحث التي تطابق في تكوينها عملية تحليل وتكشيف الوثائق. وتشتمل أيضاً على الخطوتين الأساسيتين لعملية التكشيف وهما تحليل المفاهيم، الترجمة، كما هو موضح في الشكل (2.3).

3.1.4 العلاقة بين التكشيف والاستخلاص والبحث

يوجد تداخل كبير بين هذه العمليات الثلاث (التكشيف والاستخلاص والبحث)، حيث لا يمكن فصلها في أي نظام لخزن واسترجاع المعلومات، بل إن كفاءة أي نظام لخزن واسترجاع المعلومات يعتمد على جودة هذه العمليات الثلاث. ويعد التكشيف والاستخلاص وجهين لعملة واحدة، فالتكشيف الجيد قد يستخدم في بناء المستخلصات، كما أن المستخلص الجيد يمكن الاعتماد عليه في تكشيف الوثائق. كما أن التكشيف والاستخلاص ليس لهما أي أهمية إذا لم يستخدما من أجل بحث الإنتاج الفكري وإتاحة سبل الوصول إلى أوعية المعلومات. وعلى العكس من ذلك فإن البحث دون توافر مؤشرات لمحتوى أوعية المعلومات (تكشيف واستخلاص) يجعل المستفيد مضطراً إلى أن يفحص كل وثيقة على حدة، وهو أمر غير منطقي وغير عملي في الوقت نفسه.

ويعد رضا المستفيد User Satisfaction الجانب الأساسي الذي يمكن من خلاله تقييم مدى قوة أو ضعف العلاقة بين هذه العناصر الثلاثة. فالمستفيد عادة ما يهتم بصفة عامة بالوقت المستغرق في الوصول إلى المعلومات. ولا شك أن عمليات التكشيف والاستخلاص تساعد على الوصول إلى مصادر المعلومات في أقصر وقت ممكن، حيث إنها تقدم بدائل للوثائق أكثر إيجازاً وتوفر إرشادات للوثائق الصالحة دون الحاجة إلى الرجوع إلى الوثائق الكاملة لفصل مجموعة الوثائق الصالحة عن مجموعة الوثائق غير الصالحة. كما يهتم المستفيد أيضاً بدقة النتائج المسترجعة، والتي تمثل نقطة الربط الحقيقية بين عمليات التكشيف والاستخلاص، وعمليات البحث في نظم استرجاع المعلومات.

يعتمد تحديد نوع عملية التكشيف على الطريقة التي تستخدم في الحصول على المصطلحات الكشفية، فإذا كانت المصطلحات يتم اشتقاقها من النص الأصلي يطلق عليها التكشيف الاشتقاقي Derivative Indexing أما إذا كانت المصطلحات يتم تعيينها للوثيقة فيطلق عليه التكشيف بالتعيين Assingnment Indexing. ويستخدم مصطلح التكشيف الاشتقاقي كمرادف لتكشيف الكلمات المفتاحية، نظراً لأن المصطلحات الكشفية يتم اختيارها من الكلمات الواردة بالنص مباشرة، ولا يتم الاعتماد على أي أداة لضبط المصطلحات. وعلى الجانب الآخر، فإن التكشيف بالتعيين يعتمد على

اشتقاق أو تعيين المصطلحات الدالة على مفاهيم من خلال أداة لضبط المصطلحات مثل المكانز أو قوائم رؤوس الموضوعات. وعادة ما يطلق على المصطلحات التي يتم تعيينها باستخدام المصطلحات المضبوطة الواصفات Descriptons حتى لو لم يتم تعيين تلك المصطلحات من مكنز مصطلحات. فإذا كان المفهوم الذي يتم تكشيفه جديداً أو اسم علم مثل بيت المقدس أو المسجد الأقصى ولا يجد واصفه مطابقة له بالمكنز أو قائمة المصطلحات المضبوطة، فإنه يمكن وضع محدد Identifier في عملية التكشيف بالتعيين. بمعنى آخر تحديد مصطلح جديد للدلالة على ذلك المفهوم أو اسم العلم وإضافته لأداة ضبط المصطلحات وهو ما يطلق عليه السند الأدبي في اختيار المصطلحات.

ويتم أحياناً الإشارة إلى التكشيف بالاشتقاق والذي يعتمد على أي أداة لضبط المصطلحات التكشيف الحر Free Indexing (Fugmann,1993) وتجدر الإشارة إلى أنه يوجد جدل دائر منذ بدايات النصف الثاني من القرن العشرين حول استخدام التكشيف بالتعيين أو التكشيف بالاشتقاق ومازال هذا الجدل دائراً بين المتخصصين ويمكن القول بصفة عامة إن انتشار المعلومات الرقمية أدى إلى انخفاض ملحوظ في استخدام التكشيف بالتعيين باستخدام المصطلحات المضبوطة ويرجع ذلك لعوامل تتعلق بالكم والكيف (جودة عملية التكشيف).

3.1.4.1 التكشيف الآلي والأتوماتيكي
Automated and Automatic Indexing

يتم تصنيف كل الأنشطة التي تتضمنها عملية التكشيف إلى نوعين أساسيين هما: فكري Intellectual، آلي Automated وقد تم توضيح الجزء الفكري في عملية التكشيف الذي يتضمن التحليل المفاهيمي والترجمة في الجزء السابق.

أما الجزء الآلي في عملية التكشيف فيتضمن الترتيب الهجائي وتكوين مداخل الكشاف، فبينما يتم إجراء الجزء الفكري من عملية التكشيف بالاعتماد على الجهود البشرية في معظم الأحيان، ومع التطورات المستمرة في بحوث الذكاء الاصطناعي

أصبح من الممكن إجراء عملية التكشيف بالاعتماد على الحاسبات الآلية بصورة فعالة. وتعتمد نظم التكشيف الآلي Autoamted Indexing على توظيف الحاسبات في إجراء كل من الجوانب الفكرية والميكانيكية في عملية التكشيف. ويطلق على عملية توظيف الحاسبات الآلية في إجراء الجوانب الآلية في التكشيف وقيام البشر بأداء الجوانب الفكرية مصطلح التكشيف بالآلة Automatic Indexing. من ثم فالفرق بين التكشيف الآلي والتكشيف بالآلة، أن الأول يتم كلياً بالاعتماد على الحاسبات، بينما يعتمد الثاني على إجراء الجانب الميكانيكي في تلك العملية بالاعتماد على الحاسبات فقط.

وأحياناً يُطلق على التكشيف الآلي مصطلح التكشيف الميكانيكي، حيث يُعد التكشيف الآلي أحد الحلول المبتكرة لمشكلات عدم الاطراد Inconsistemcy والكُلفة المرتفعة المرتبطة بالتكشيف اليدوي. مع ذلك فإن نقطة الضعف الجوهرية في التكشيف الآلي تتمثل في أنه يتعامل مع الجانب الفكري في عملية التكشيف بكفاءة أقل بكثير من إمكانيات أخصائيي المعلومات المحترفين. ويرجع السبب في ذلك إلى أن الحاسبات لا تستطيع التفكير ولا تملك القدرات التحليلية للبشر. وفي المقابل، يحرر التكشيف الآلي المكشفين المحترفين من بعض المهام التكشيفية التكرارية المملة، من ثم يمكنهم التركيز على العمليات الفكرية للتكشيف. وتزداد قيمة التكشيف الآلي بصورة أكبر مع تضخم المعلومات المتاحة فى البيئة الرقمية والتي تنمو بمعدلات كبيرة تتجاوز ملايين الجيجابايت التي تنتج يومياً في البيئة الرقمية. ويعتمد التكشيف الآلي على العديد من الأساليب التي تم تطبيقها بالاعتماد على خوارزميات تردد المصطلحات Term Frequency، تقارب المصطلحات Keyword Poroximity، مواضع المصطلحات Term Locations، التكشيف الاحتمالي Probability Indexing، واللغويات Linguistics. وقد تم توظيف المصطلحات المضبوطة في بعض إجراءات التكشيف الآلي، ولكنها لم تحقق النجاح المطلوب وأثرت سلباً في الطبيعة الحاسوبية لذلك النشاط.

3.1.4.2 التكشيف في بيئة الروابط الفائقة

Indexing in the hyper text Environment

ينمو حجم المعلومات المتاحة في بيئة الروابط الفائقة بسرعة كبيرة، وترمز تلك البيئة إلى الشبكة العنكبوتية العالمية أو شبكة الويب، وتعتمد المعلومات المتاحة على الويب في تمثيلها لمصطلحات الكشاف على استخدام الروابط الفائقة، والتي تجسد كلاً من مصطلحات الكشاف وآلية تحديد موقع المعلومات.

وبمعنى آخر يتم توظيف الروابط الفائقة على أنها مصطلحات كشفية Indexing Terms، حيث تقود تلك الروابط الفائقة المستفيد بسلامة إلى المواقع التي تشير إليها مصطلحات الكشاف.

وبالمقارنة مع غيرها من بيئات التكشيف فإن هذه البيئة تتميز بالملامح التالية:

أولاً: مصطلحات الكشاف في بيئة الروابط الفائقة تمثل جزءاً أصيلاً من الوثائق التي يتم تكشيفها وليست كيانات مستقلة خارج النص الذي يتم تكشيفه.

ثانياً: يتم دمج مصطلحات الكشاف، وموضوعات الوثائق معاً في وحدة واحدة بدلاً من فصلها في قوائم مستقلة.

ثالثاً: من الصعب التعرف في تلك البيئة إلى البنية الهرمية للموضوعات والمفاهيم الفرعية وعلاقاتها ببعضها بعضاً، كما هو الحال في البيئة التقليدية للتكشيف.

رابعاً: يمكن فقط في تلك البيئة استخدام الروابط الفائقة التي تحتوي على مؤشرات محتوى Content Base ID Link كمصطلحات كشاف، ومن ثم لا يتم توظيف الروابط التنظيمية Organizational Links مثل الصفحة التالية، السابقة، أعلى الصفحة في عملية التكشيف (Chu & Rosenthal, 1995).

خامساً: يهتم القائمون على إعداد الوثائق التي يتم إتاحتها في بيئة الربط الفائق بدور عملية التكشيف التي تتم أحياناً بالتزامن مع عملية بناء الوثيقة وأحياناً قبلها. وكنتيجة لذلك فإن مصطلحات مثل (انقر هنا) والتي نادراً

ما يتم اختيارها كمصطلحات كشفية من جانب المكشفين تظهر في هذه البيئة على أنها أسماء لروابط فائقة من ثم يتم تكشيفها.

سادساً: تقلل تلك البيئة التضارب الذي يحدث بين الوثيقة الأصلية والمصطلحات الكشفية؛ حيث يقرر منتج الوثيقة عند بنائها من البداية ما هي المصطلحات التي تستخدم في وصف الروابط الفائقة من ثم يتم تكشيفها، أما الوثائق التقليدية فيتم كتابتها أولاً ثم يقوم المكشف بتحليل الوثيقة واختيار المصطلحات الكشفية بغرض تمثيلها.

وبناءً على الملامح الخاصة بعملية التكشيف في بيئة الروابط الفائقة، يجب استخدام الطرق الملائمة في تكشيف تلك الوثائق. فعلى سبيل المثال يجب اختيار أسماء الروابط بعناية عند إعداد وثيقة يتم نشرها عبر بيئة الروابط الفائقة، لذلك ظهر مجال مهم في تكشيف تلك البيئة يطلق عليه تحسين أداء محركات البحث Search Engines Optimization.

3.2 التوسيم الاجتماعي
Social Tagging

ظهر التوسيم الاجتماعي مع بدايات الجيل الثاني للويب الذي تحول فيه المستفيد في بيئة العنكبوتية من مُستقبل للخدمة إلى مُتفاعل مع الشكبة، ثم تطور بصورة أكبر مع التوسع في بيئة الويب الدلالي التي تركز على الربط بين الخدمات وإبراز المعاني والدلالات التي تحملها الصفحات. ويتم من خلال أدوات التوسيم الاجتماعي تجميع كلمات مفتاحية من المستفيدين من مصادر الويب على منصة تستخدم في وصف الكيانات والمفاهيم والأفكار التي تحملها تلك المصادر.

ومن المعروف أنه توجد أنماط متعددة للتوسيم استخدمت في المكتبات منذ القدم، منها استخدام الملصقات Labels والتي تطورت إلى الأكواد العمودية Barcode أو محددات الهوية بترددات الراديو[1] (RFID). ومع تطور أساليب التواصل الاجتماعي ظهر

(1) RFID: Radio-Frequency IDentification

التوسيم الاجتماعي كآلية جديدة مختلفة عن تلك الأشكال التقليدية التي استخدمت في تحديد هوية الوثائق. وقد ظهر التوسيم الاجتماعي في بداية عام 2003 كوسيلة يستخدمها المستفيدون في إثراء المصطلحات الدالة على الوثائق المتاحة على الإنترنت، فيما عُرف بالتكشيف الاجتماعي Social Indexing. بالتالي فالتوسيم الاجتماعي يعد أحد الأنشطة التي يمارس فيها المستفيد النهائي عملية التكشيف بالكلمات المفتاحية، وتتم عملية التكشيف هنا بعد إتاحة الوثيقة للمستفيد الذي يقوم بتكشيفها أو تتم بطريقة آلية من خلال نظام استرجاع المعلومات الذي يُخزن نتائج تفاعل المستفيد مع النظام. من ثم فإن التوسيم الاجتماعي ليس مساوياً تماماً أو مطابقاً للتكشيف بالكلمات المفتاحية، نظراً لأن المستفيد عندما يقوم بعملية التوسيم يختار أسماء أو عبارات تستخدم للدلالة على الوثيقة أو لوسم (تسمية) الوثيقة وليس تكشيفها.

ويعد التوسيم الاجتماعي أحد أنماط حركة الجيل الثاني للويب التي تسعى إلى توسيع نطاق مشاركة المستفيد في بث وإتاحة المعلومات مثل المدونات Blogging والويكيبيديا.. الخ. ويُعد موقع فيلكر Flicker لمشاركة الصور من أقدم أنظمة التوسيم، كما يُعد موقع del.icio.us الذي تغير عنوانه إلى delicious.com في عام 2007 أيضاً من أقدم أنواع هذه النوعية من المواقع. ففي مثل هذه النوعية من المواقع يستطيع المستفيدون التعليق على الوثائق النصية أو الوسائط المتعددة المتاحة على الويب بكلمات أو عبارات من اختيارهم يمكن أن تستخدم في بحث واسترجاع تلك الوثائق.

وتوجد العديد من الأدوات التي تتيح للمستفيد إضافة كلمات مفتاحية للوثائق التي تكشفها أدوات بحث والاسترجاع على الإنترنت، لعل أبرزها محرك البحث Pubmed والذي يعد أحد أهم وأبرز قواعد البيانات الطبية والذي يصدر عن المكتبة القومية الطبية ويهتم بتمثيل وتكشيف مصادر المعلومات الطبية من درويات وأعمال مؤتمرات.. إلخ، حيث يتيح للمستفيد التوسيم الاجتماعي للوثائق وينتج عنها سحابة الواسمات Tag Cloud.

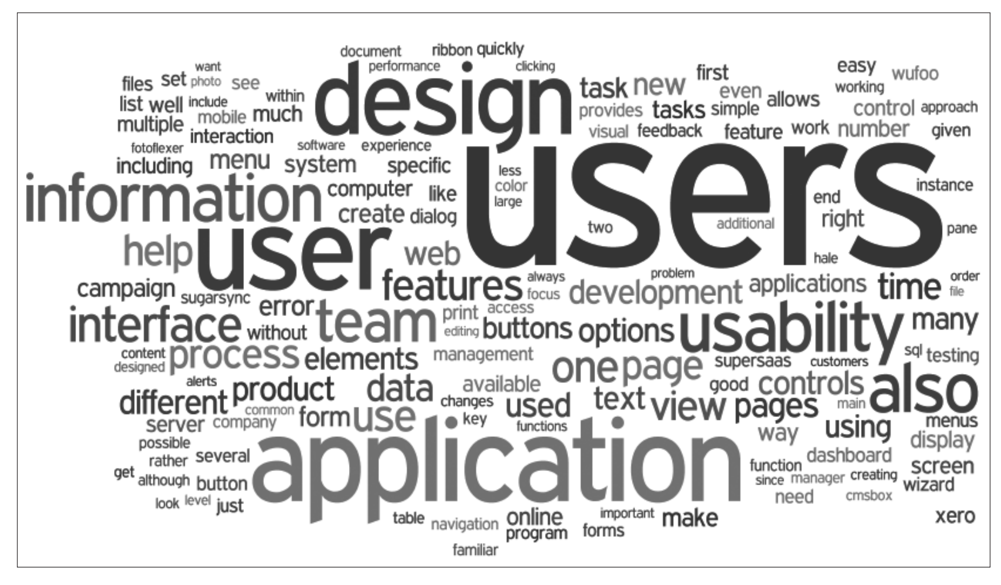
على الرغم من أن التوسيم الاجتماعي قد فتح مجالاً جديداً في تمثيل واسترجاع المعلومات يتيح للمستفيد إمكانيات المشاركة الفعالة في عمليات التمثيل، فإنه يعاني من نفس المشكلات التي تظهر في التكشيف الآلي مثل القصور الذي يبرز في

عمليات التكشيف بالكلمات المفتاحية ومنها المترادفات والمشترك اللفظي.. الخ، والتي تظهر بوضوح وعلى نطاق واسع في التوسيم الاجتماعي.

وعلى الرغم من ذلك فإن التوسيم الاجتماعي يُعد نمطاً متميزاً وأحد البدائل المهمة التي أتاحتها بيئة الشبكة العنكبوتية لتمثيل المعلومات وتيسير استرجاعها، نظراً لأن الواسمات التي يضعها المستفيدون، إضافة إلى مزاياها الأخرى، تتيح نقاط إتاحة إضافية يتم اختيارها من جانب المستفيد النهائي كمصطلحات استفسار لتيسير الوصول إلى المعلومات، وتمكن المستفيدين الآخرين من التوسع في البحث وفهم النتائج المسترجعة من خلال الواسمات المستخدمة.

وقد ساعد التوسيم الاجتماعي كأحد الأنماط الجديدة في تمثيل المعلومات على ابتكار أساليب لإثراء مجال استرجاع المعلومات (Smith,2008).

ويعد التقسيم الجماعي Folksonomies أحد أبرز تلك الابتكارات، ويشير مصطلح التقسيم الجماعي، الذي سكّه لأول مرة العالم توماس فاندر Thomas Vander في



نموذج (1)

نموذج لسحابة كلمات من موقع world.net

عـام 2004 إلـى مكونيـن أساسيين همـا المجتمـع Folks والتقسـيم Taxonomy وبعبـارة أخـرى فـإن التصنيـف الاجتماعـي هـو عبـارة عـن نظـام تصنيـف تـم بنـاؤه باسـتخدام واسمات Tags أنشأها المجتمع أو المستفيدون النهائيون، وسوف يتم مناقشة التصنيف الاجتماعـي والوسـم الاجتماعـي فيمـا يلـي.

وعـادة مـا يأخـذ الوسـم الاجتماعـي شـكل سـحابة الواسـمات Tags Cloud والتـي تعـد تجميعـاً مرئيـاً للواسـمات Visual Alggregation of Tags يتـم عرضهـا فـي مواقـع الوسـم Tagging sites بالاعتمـاد علـى تـردد الوسـم Tagged Frequencies وتسـاعد سـحابة الواسـمات المسـتفيدين علـى اختيـار المصطلحـات الملائمـة سـواء فـي عمليـة الوسـم أو الاسـترجاع.

3.3 التقسيم إلى فئات

Categorization

يسـاعد التقسـيم إلـى فئـات علـى تمثيـل المعلومـات بصـورة هرميـة متتاليـة فـي البنـاء توضـح الأقسـام والأجـزاء التـي ينتمـي إليهـا كل قسـم. وينقسـم هـذا النـوع مـن أنـواع تمثيل المعلومـات إلـى نمطيـن أساسيين، سيتم منقاشـتهما هنـا بالتفصيـل فـي القسـم التالـي.

3.3.1 أنماط التقسيم إلى فئات

Types of Categoration

يعتمـد النمـط التقليـدي للتقسيم إلى فئات على استخدام نظم تصنيف المعرفة التقليدية مثـل خطـة تصنيـف ديـوي العشـري، مكتبـة الكونجـرس. ويطلـق علـى هـذا النمـط مـن أنمـاط التقسـيم إلـى فئـات عالميـاً مصطلـح التصنيـف Classification والـذي يتـم تطبيقـه بصفـة عامـة علـى مقتنيـات المكتبـات وخدمـات المعلومـات، ويعتمـد التصنيـف علـى اسـتخدام أسـاليب متنوعـة لترميـز المعلومـات تشـمل الأرقـام والحـروف أو مزيجـاً منهمـا إلـى جانـب العلامـات الخاصـة.

ومـع تقـدم الإنترنـت وانتشـار اسـتخدامها فـي بـث وإتاحـة المعلومات مـن خـلال مواقـع

الويب، اتخذت المعلومات التي يتم بثها من خلال هذه البيئة مجموعة من الملامح الجديدة تشمل المعلومات العابرة التي يتم إزالتها أو تغييرها وتعديلها بسرعة، ونظراً لأن المعلومات المتاحة مختلفة في مدى جودتها (حيث إنه لا يوجد أي رقابة عليها) إلى جانب ضخامة الحجم. لذلك فإن استخدام نظم التصنيف التقليدية في تمثيل المعلومات لتقسيم هذا الكم الهائل سريع التغيير والمتنوع في مدى جودته يبدو أمراً مكلفاً للغاية، وغير ملائم لطبيعة تلك المعلومات. ومن هنا ظهرت الحاجة إلى نظام جديد لتقسيم المعلومات المتاحة على الإنترنت إلى فئات فظهر تصنيف الويب Web Taxonomy والذي يعتمد على استخدام فئات واسعة لتقسيم مواقع وصفحات الويب. ويُعد دليل البحث Yahoo الأداة الرائدة في هذا المجال، والذي أصبح فيما بعد أحد أبرز نماذج تمثيل المعلومات على الويب.

وتعتمد نظم تصنيف الويب على تقسيم المواقع والصفحات إلى فئات واسعة ثم أقسام أكثر تخصيصاً مع وضع روابط فائقة مباشرة تغني عن استخدام نظم الترميز الرقمي أو الهجائي والتي تعكس إطار البناء الهرمي والعلاقات بين الفئات.

3.3.2 مبادئ التقسيم إلى فئات

عند استخدام التقسيم إلى فئات لتمثيل المعلومات يتم التعبير عن الوثيقة بفئة واحدة وأحياناً اثنتان وذلك في حالة المواد التي تعالج موضوعات متداخلة. وبمعنى آخر يتم تصنيف كل وثيقة تحت فئة واحدة فقط من الفئات المحدده بنظام التقسيم.

وتتطلب هذه الممارسة أن تكون الفئات المختارة بنظام التقسيم إلى فئات تتميز بما يلي:

- الشمولية Exhaustive
- الحصرية التبادلية Mutually Exclusive

من ثم يمكن القول إن نظام التقسيم إلى فئات لا بد أن يشتمل على كل الفئات المحتملة لتمثيل المعلومات بدقة. وفي الوقت نفسه، يجب أن تكون هذه الفئات حصرية

بشكل تبادلي وواضح (بمعنى أنه يمكن تكرارها). فإذا كان النظام لا يحقق الملمح الأول، فإن بعض المعلومات سيكون من الصعب تمثيلها وفقاً للفئات المتاحة بنظام التقسيم. وإذا لم يتحقق الملمح الثاني يكون من الممكن استخدام أكثر من فئة واحدة لتمثيل الموضوع نفسه في نفس الوثيقة. كما أن عدم توافر أي منهما أو كليهما يضعف تماسك نظام التقسيم إلى فئات. ومن المبادئ المهمة أيضاً التي يجب أن تتوافر في أي نظام للتقسيم إلى فئات: المرونة وسهولة الاستخدام ولكنها ليست مبادئ أساسية.

وقد سعت معظم أدوات الوصول إلى المعلومات على الويب إلى بناء أدلة بحث تعتمد على تقسيم الويب إلى فئات مع بدايات ظهور محركات البحث في عام 1994 ومنها دليل البحث ياهو Yahoo.com ودليل البحث جوجل. وقد قسم كل منهما الويب إلى 14 فئة موضوعية أساسية وتحت كل فئة رئيسة مجموعة من الفئات الموضوعية الفرعية التي وصلت إلى أكثر من 90 فئة فرعية. وتجدر الإشارة إلى أن أدلة ياهو وجوجل تم إغلاقها منذ عام 2014. ولعل أبرز الأمثلة للتقسيم إلى فئات في قواعد البيانات هو إمكانيات التصفح التي تتيحها الكثير من قواعد البيانات الدولية لعل أبرزها قاعدتا بيانات Web of Science وScopus. ويمكن مراجعة الفئات الموضوعية لقاعدة بيانات Scopus من خلال مراجعة الموقع الخاص بتقرير Scimago المتاح على الرابط التالي: https://www.scimagojr.com/journalrank.php.

3.3.3 العلاقة التي تجمع بين الاتجاهين

تتشابه الطريقتان المستخدمتان في التقسيم إلى فئات في العديد من الملامح، لعل أبرزها هو تمثيل المعلومات في صورة فئات لها بنية هرمية تعتمد على قوة العلاقة بين مصدر المعلومة والفئة التي ينتمي إليها، كما أن الفئات عادة ما تلتزم تتابع خطي في عمليات البناء والوصول إلى المعلومات. ونظراً لعدم قدرة نظم التصنيف التقليدية على متابعة التطورات المتسارعة في حجم الويب وطبيعتها الترابطية، ظهرت نظم تصنيف الويب التي أطلق عليها أدلة البحث في البداية، ثم تطورت تلك النظم إلى أدوات تعتمد على أساليب التنقيب عن البيانات Data Mining وعناقيد الويب

Web Clustering والتي تستخدم أساليب التحليل الدلالي للمفاهيم بالاعتماد على نظم تصنيف الويب أو التوكسونومي. مع ذلك توجد بعض الاختلافات الأساسية بين الاتجاهين، وتعتمد هذه الاختلافات على طبيعة الإطار المستخدم لأغراض تمثيل المعلومات. فقد تم استخدام التصنيف مع أنواع متعددة ومتنوعة من مصادر المعلومات، وأثبت تميزاً كأحد أساليب تمثيل المعلومات، أما تصنيف الويب فعادة ما ينظر إليه على أنه طريقة سريعة ومرنة في تمثيل المعلومات. ومع ازدياد حجم المعلومات الثابتة التي أصبحت ذات أهمية كبيرة بالنسبة إلى المستفيدين من الويب، بدأ استخدام التصنيف التقليدي في تقسيم المعلومات المتشابكة على العنكبوتية، وفي الوقت نفسه تحسنت النماذج المستخدمة في بناء نظم تصنيف الويب من خلال تطبيق نماذج معمارية الويب Web Archetitutcre والتي نشأت أساساً اعتماداً على نظم التصنيف التقليدية مثل التمثيل الهرمي.

علاوة على ذلك، فإن تقسيم النصوص إلى فئات، من وجهة نظر تقنيات الميكنة، ينطبق بصورة أكبر على تصنيف الويب أكثر من التصنيف التقليدي، حيث إن حجم المعلومات الرقمية يتزايد بسرعة كبيرة. فمع اهتمام الباحثين بالتصنيف الآلي Automatic Classification حدث تقدم كبير في آليات التقسيم إلى فئات، إلا أنه توجد حاجة ماسة إلى توظيف العقول البشرية للخروج بنظم تصنيف دقيقة، والتي يتعذر تحقيقها مع استخدام خوارزميات تعتمد على الآلات فقط. بمعنى أن التدخل البشري عنصر مهم في تلك العملية حتى الآن.

3.3.4 التلخيص Summarization

التلخيص هو تعبير موجز ومختصر للمحتوى المعلوماتي، بحيث يصف ذلك الحقائق والأفكار الأساسية التي تتضمنها الوثيقة. وتوجد أربع طرق أساسية في التخليص في البيئة الرقمية هي المستخلصات والملخصات والاشتقاقات والتلخيص الوافي للموقع، ولكل طريقة من هذه الطرق أدواتها وآلياتها. وسيتم فيما يلي استعراض تلك الطرق ومخرجات كل منها:

3.3.4.1 المستخلصات Abstracts

المستخلص عبارة عن تمثيل مركز ودقيق لمحتوى الوثيقة بالاعتماد على أسلوب إعداد المستخلصات والذي يتم تنفيذه من خلال أخصائيين مؤهلين لأداء تلك العملية، ذلك على الرغم من محاولة تطوير أساليب آلية في الماضي (Luhn,1958). ويجب أن يتم كتابة المستخلص بأسلوب يشبه بدرجة كبيرة الوثيقة الأصلية، على الرغم من صعوبة تحقيق هذا المبدأ أثناء عملية التلخيص، بسبب الحاجة إلى حذف كثير من المعلومات أثناء عملية إعداد المستخلص، ما يؤدي إلى قصور في تمثيل المستخلص للوثيقة. ويتم تقسيم المستخلصات إلى ثلاثة أنواع هي:

- **المستخلصات الإعلامية Informative Abstracts**
- **المستخلصات الدلالية Indicative Abstracts**
- **المستخلصات النقدية Critical Abstracts**

المستخلصات الإعلامية تحتوي على المعلومات الأساسية التي تعالجها الوثيقة الأصلية، لذلك من الممكن أن تستخدم كبديل للوثيقة في بعض الأحيان. وبناء على المستخلص الإعلامي يمكن أن يقرر الباحث ما إذا كان في حاجة إلى قراءة الوثيقة الأصلية أم لا.

أما المستخلصات الدلالية فهي وصف موجز للمحتوى المعرفي الذي تتضمنه الوثيقة Aboutness، مع استبعاد التفاصيل مثل المنهج والنتائج. لذلك لا يمكن معاملة المستخلصات الدلالية على أنها بديل للوثيقة الأصلية، حيث يحتاج الباحث إلى الرجوع إلى الوثيقة الأصلية للحصول على التفاصيل التي لا تتضمنها المستخلصات الدلالية.

المستخلصات النقدية لا تقتصر فقط على تمثيل المعلومات التي تشتمل عليها الوثائق، ولكنها تحاول أيضاً تقييم تلك المعلومات والحكم على جودتها وصلاحيتها. وقد بدأت العديد من قواعد بيانات الأدلة والبراهين Evidance Based Databases الاعتماد بكثافة على هذه النوعية من المستخلصات من خلال خبراء يقومون بكتابة مراجعات نقدية عن الأبحاث في صورة ملخصات وانتقاء أفضل النتائج التي

توصلت إليها الدراسات ووضعها في قواعد بيانات جديدة يطلق عليها قواعد بيانات الأدلة والبراهين، والتي يعتمد الكثير منها على إعادة التجارب في مختبرات معتمدة والتعليق النقدي على البحوث ومقارنتها بنتائج المختبرات.

ويختلف هذا النمط من أنماط الاستخلاص عن المغزى الأساسي من الاستخلاص الذي يجب أن يكون موضوعياً ومجرداً من أي تفسيرات إلا نادراً أو من جانب فئات تمتلك القدرة على الحكم النقدي في المجالات العلمية. لذلك لا يقوم أخصائي المعلومات بكتابة مستخلص نقدي بصفة عامة إلا إذا طُلب منه ذلك.

وكما ذكر سابقاً، قام العديد من الباحثين بمحاولات لإنتاج برامج للاستخلاص الآلي؛ إلا أن المنتج النهائي لتلك المحاولات لا يختلف كثيراً عن التلخيص الآلي أو الاشتقاق الآلي، أكثر من كونها استخلاصاً آلياً Auto Abstract حيث تشتمل على مجموعة من الجمل المفتاحية التي يتم اشتقاقها من الوثيقة الأصلية.

3.3.4.2 التلخيص Summaries

هو عبارة عن إعادة صياغة لمجموعة النقاط الرئيسة التي تعالجها الوثيقة الأصلية، ويتم وضع الملخص إما في بداية الوثيقة أو في نهايتها. وعلى الرغم من التشابه الكبير بين الملخص والوثيقة الأصلية، إلا أنه لا يغني عن الوثيقة الأصلية، حيث يفترض معد هذه النوعية من الملخصات أن القارئ سوف يتابع قراءة الوثيقة الكاملة، لأن هذا النمط عادة ما يفتقر إلى العناصر الأساسية اللازمة لفهم الوثيقة مثل الأجزاء الخاصة بالمعلومات المتعلقة والخلفيات المعرفية للموضوع والمنهج وآليات الوصول إلى النتائج.. الخ (Rowley,1994).

وقد تم في السنوات الأخيرة تطوير العديد من خوارزميات التلخيص الآلي للنصوص وخاصة النصوص الرقمية (Jones, 2007) ويعد التلخيص الآلي أحد المجالات النشطة التي يهتم بها الباحثون في مجالات الذكاء الاصطناعي ومعالجة اللغة الطبيعية. وقد أطلق بعض الباحثين على المخرجات التي تنتجها خوارزمياتهم

مصطلح مستخلصات Abstracts؛ إلا أنها لا تُعد ملخصات آلية للوثائق الأصلية، ومع ذلك فإن أنظمة الذكاء الاصطناعي هي الوحيدة القادرة على تحويل حلم الاستخلاص الآلي إلى حقيقة يمكن إنجازها وهذه الخوارزميات لم يتم إنجازها بنجاح إلى الآن. ومن أهم العقبات التي تواجه إنتاج مستخلصات آلية، معالجة الدلالات وفهم النصوص Semantic Porcessing and Text Understanding من خلال أنظمة التلخيص الآلي.

◄ 3.3.4.3 الاشتقاقات Extacts

الاشتقاق هو عبارة عن جزء أو أكثر من الوثيقة يتم اختياره لتمثيل الوثيقة ككل، ولا يمكن لتلك الاشتقاقات أن تمثل الوثيقة بشكل جيد؛ مع ذلك فهي مفيدة للقارئ الذي يحتاج إلى موجز لأغراض دراسة معينة، ولا يمكن النظر إلى الاشتقاق بأي حال من الأحوال على أنه بديل للوثيقة الأصلية. على الرغم من أنه عادة ما يتم النظر إلى الاشتقاق على أنه أقل من حيث الكفاءة وجودة التمثيل عن كل من الاستخلاص والتلخيص؛ إلا أنه يعتمد بصورة كاملة على النظم الآلية. فجميع أنظمة استرجاع المعلومات على الإنترنت بما فيها جوجل تعتمد كلياً على الطرق الآلية للاشتقاق.

ومن الأساليب التي تم استخدامها من جانب نظم الاسترجاع على الإنترنت في الاشتقاق هو استخدام نموذج القطع Ellipsis أو التوقف عند نقطة معينة عن إجراء الاشتقاق عندما يصل الجزء المشتق إلى نقطة القطع Cut off Point التي تم تحديدها بخوارزميات النظام. لذلك فإن جودة عملية الاشتقاق الآلي تعد إحدى المشكلات المهمة التي يتم النظر إليها في بحوث ودراسات هذه النوعية من النظم.

◄ 3.3.5 الملخص الوافي للموقع (موم)

يمكن وضع مختصرة عربية موازية للمصطلح RSS وهي (موم) لتشير إلى مصطلح الملخص الوافي للموقع، والذي يُعد أحد تطبيقات الجيل الثاني للويب ويستخدم لأغراض تمثيل المعلومات بصورة موجزة ومختصرة.

وبشكل أكثر تحديداً يتم استخدام موم RSS مع أشكال الملفات التي يُطلق عليها التجميع لأغراض التغذية للمعلومات المحدثة من مصادر متنوعة. ويمكن للأفراد المشتركين في هذه النوعية من الخدمات من خلال قارئ يطلق عليه برنامج التجميع Aggregator Program أن يستقبلوا على أجهزتهم الخاصة المعلومات المحدثة التي تتيحها برامج التغذية، لذلك يمكن النظر إلى موم على أنها خدمة إحاطة جارية في بيئة الويب، تقدم للمشتركين فيها ملخصاً للمعلومات الحديثة المتاحة من المصادر التي يهتمون بها.

وإدراكاً للدور المهم لخدمة موم قام المطورون في اتحاد الشبكة العنكبوتية العالمية W3C بتطوير إصدار جديد من موم، عندما توقفت شركة Netscape والتي طورت أول متصفح ويب بالرسومات عن دعم الإصدار الأول من قارئ موم الذي قامت بتطويره. نظراً لأن الإصدار الجديد من موم تم بناؤه بالاعتماد على معيار إطار وصف المصادر Resources Description Framework- RDF والذي قامت W3C أيضاً بتطويره كجزء من حركة الويب الدلالي التي يدعمها الاتحاد، فقد تم تغيير استهلالية موم لتصبح RDF Site Summary أي ملخص الموقع باستخدام إطار وصف المصادر، وذلك لتمييزه عن الإصدار السابق (Kelly,2005). وتجدر الإشار إلى أنه يوجد مصطلح آخر مستخدم للدلالة على مفهوم موم وهو التلقيم الحقيقي المبسط – Really simple syndication RSS ـ والذي يعتمد على التقنية والأدوات نفسها.

```
- <item>
  - <title>
      <![CDATA[ Countdown for nasty Windows virus  ]]>
    </title>
    <link>http://news.bbc.co.uk/go/rss/-/2/hi/technology/4661582.stm</link>
  - <description>
      <![CDATA[ A destructive Windows virus is set to start deleting
      popular files on infected machines on 3 February.  ]]>
    </description>
  - <author>
      <![CDATA[ boris@yeltsin.com(Boris Yeltsin)  ]]>
    </author>
  - <category domain="http://www.MyDomain/technology">
      <![CDATA[ Technology  ]]>
    </category>
    <comments>http://news.bbc.co.uk/go/rss/-
      /2/hi/technology/4661582.stm</comments>
    <enclosure url="http://news.bbc.co.uk/go/rss/-
      /2/hi/technology/fake_video_link.mpeg" length="99554122"
```

نموذج (2) لشكل ملف موم RSS XML FORMAT

وعند مقارنة موم مع غيره من طرق التلخيص التي تم تناولها في هذا الجزء، نجد أن موم يتم بطريقة آلية على الويب. وتلبي هذه الطريقة الآلية احتياجات قطاع عريض من المستفيدين على الويب الذين يرغبون في الحصول على المعلومات الحديثة التي تظهر في مجموعة من المواقع في مكان واحد. فقارئ الملخص الوافي للمواقع أو التلقيم المبسط للمحتوى يقوم بتجميع Aggragate المعلومات الموجزة من مناطق معينة في مواقع الويب وعرضها للمستفيد في مكان محدد بموقعه.

وتجدر الإشارة إلى أن جودة الملخص الذي تنتجه هذه الطريقة أقل بكثير من غيرها من طرق التلخيص مثل المستخلصات، حيث إن جودة عملية التمثيل لا تستند إلى معايير محددة في إعداد الملخص الوافي للموقع، ما يجعلها متضاربة في الشكل ومختلفة في البناء على عكس المستخلصات التي توجد معايير تحدد طريقة إعدادها وأشكال البناء الخاصة بها.

ونظراً لأن التكشيف يعد أبرز نماذج تمثيل المعرفة وأكثرها استخداماً في البيئة الورقية والرقميةأيضاً، فمن الضروري تسليط الضوء على أنواع الكشافات وطرق تقسيمها وبنائها ووظيفة كل منها كأدوات لتمثيل المعرفة.

3.4 أنواع الكشافات

يرى عبدالهادي (2005) أنه يمكن تقسيم الكشافات بناءً على طبيعة الوحدات المكشفة، نوعية المداخل المستخدمة، طريقة ترتيب المداخل، نضيف إلى ذلك أنه يمكن النظر إلى الكشافات أيضاً وفقاً لنظام التكشيف المستخدم إلى كشافات آلية وكشافات مميكنة وكشافات يدوية كما سبق وأوضحنا.

3.4.1 تقسيم الكشافات وفقاً لطبيعة المادة المكشفة

تنقسم الكشافات وفقاً لطبيعة المادة المكشفة إلى خمسة أنواع أساسية هي:

◄ 3.4.1.1 كشافات الكتب

Books Index

يتم في تلك النوعية تكشيف المفاهيم والأعلام والمصطلحات الواردة في نصوص الكتب، وتلحق بنهايات الكتب، لكي تستخدم في الوصول إلى أي معلومة تفصيلية بالكتاب عند الحاجة. وعادة ما يتم ترتيبها ترتيباً هجائياً منفصلاً لكل نوعية بحيث يكون لكل شكل كشاف منفصل (كشاف للأعلام، آخر للأماكن، ثالث للمفاهيم أو الكلمات المفتاحية)؛ أو ترتيباً شاملاً يجمع كل هذه العناصر مجتمعة معاً في كشاف واحد. ويستخدم هذا النوع من الكشافات في الكتب كما يستخدم أيضاً وعلى نطاق واسع في معظم أنواع المواد المرجعية مثل الموسوعات، الكتب السنوية، الأدلة.. إلخ.

◄ 3.4.1.2 كشافات المسلسلات

Serials Index

هي عبارة عن كشافات بمحتويات الدوريات والصحف والمجلات من مقالات وأخبار وغيرها. وغالباً ما ترتب هذه الكشافات ترتيباً هجائياً واحداً. ويعد هذا النوع من الكشافات من أكثر الأنواع شيوعاً وأهمية، نظراً لما مر به من تطورات بدأت باستخدام الحاسب الآلي في عمليات التكشيف، والبحث خارج الخط المباشر ثم البحث على الخط المباشر وأخيراً الاسترجاع من خلال شبكة الإنترنت والشبكة العنكبوتية.

◄ 3.4.1.3 كشافات الاستشهادات المرجعية

Citations Indexes

إذا كانت كشافات الدوريات تساعد على الوصول إلى مقالات الدوريات التي تم تكشيفها تحت رؤوس موضوعات أو كلمات مفتاحية تصف محتواها الموضوعي، فإن كشافات الاستشهادات المرجعية تساعد على الوصول إلى مقالات الدوريات وفقاً للعلاقات التي تربط بينها من خلال الاستشهادات المرجعية. فالعلاقة التي تتشكل بين المقالة المصدرية والأعمال التي تم

الاستشهاد بها في هذه المقالة المصدرية تعني وجود رابطة خفية بين المفاهيم والموضوعات التي تمت معالجتها في المقالة المصدرية والأعمال المستشهد بها، وهو الأساس الذي تقوم عليه فكرة كشافات الاستشهادات المرجعية. فقد استقى يوجين جارفين Ugine Garven فكرة كشافات الاستشهادات المرجعية من فكرة السوابق القانونية المستخدمة في القانون الأمريكي. وتساعد هذه الكشافات على التعرف إلى الدوريات البؤرية، الأعمال البؤرية في تخصص ما، والمؤلفين البؤريين أو الأساسيين في أحد المجالات العلمية. فتكرار الاستشهاد بمؤلف معين في أحد المجالات يعني أن دراسات هذا المؤلف من الأعمال البؤرية في ذلك المجال الموضوعي. وسيتم عرض نماذج لتلك النوعية من الكشافات عند استعراض قضية التمثيل في نهاية هذا الفصل.

◄ 3.4.1.4 كشافات النصوص

Concordance Indexes

تتيح تلك النوعية من الكشافات تحليلات صرفية كاملة للمواد ذات الطبيعة الخاصة بحيث يمكن الوصول إلى كل جذور الكلمات ومشتقاتها في تلك النصوص. وعادة ما تستخدم هذه الآلية في تكشيف النصوص المهمة مثل النصوص الدينية والكتب المقدسة والقوانين والدساتير والاتفاقيات والمعاهدات والأعمال الأدبية البارزة.. الخ. وعادة ما ترتب هذه الكشافات هجائياً وفقاً للمصطلحات الواردة في النصوص متبوعة بأماكن وجودها في متن النص. ويتم إعداد هذه الكشافات لكل كلمات النص دون تمييز. يستخدم هذا النوع من الكشافات مع النصوص ذات القيمة الكبيرة، ويكون لكل كلمة في النص أهمية لا يمكن إغفالها. ومن أمثلة هذا النوع من الكشافات «المعجم المفهرس لألفاظ القرآن الكريم/ محمد فؤاد عبد الباقي»، و«المعجم المفهرس لألفاظ الحديث إعداد فنسنك، أي، تحقيق محمد فؤاد عبد الباقي».

ويتميز هذا النوع من الكشافات بإمكانية البحث فيه بأي كلمة في النص، ما يساعد على تحديد موضعها أو بيان موقعها ضمن جملة أو سياق معين. ويستخدم

أيضاً في الدراسات اللغوية والمعجمية حيث إن العديد من التفاسير اللغوية تعتمد على مثل هذا النوع من الكشافات في تجميع المعاني المختلفة لمفهوم واحد. ويعد هذا النوع من الكشافات من أصعب أنواع الكشافات في حالة النظم اليدوية، لكنه يعد من أسهل وأسرع أنواع الكشافات في حالة نظم التكشيف الآلي التي تعتمد على استخدام إمكانيات الحاسب الآلي في تحديد مواضع الكلمات والجمل. فعلى سبيل المثال في حالة استخدام هذا النوع من الكشافات في تحديد عدد مرات ورود كلمة الجنة والنار في القرآن الكريم، ثم تحديد مواضع ورودهما سواء معاً أو بشكل منفصل. يقوم نظام التكشيف الآلي بإعراب Parsing للنص بالكامل بحثاً عن الكلمتين باستخدام أسلوب المضاهاة المضبوطة Exact Match – أي مضاهاة حرف بحرف – وعندما تتطابق كل الحروف مع بعضها بعضاً يعرض نظام التكشيف الكلمة مصحوبة بالسياق مثل السورة ورقم الآية وغيرها من المحددات التي يمكن التحكم فيها قبل إجراء البحث.

◄ 3.4.1.5 كشافات مواقع الإنترنت
Internet Indexes

يطلق على هذه النوعية من الكشافات أدوات تمثيل واسترجاع المعلومات المتاحة على الإنترنت. يوجد أربع أدوات رئيسة يمكن استخدامها في بحث الشبكة العنكبوتية هي أدلة البحث ومحركات البحث، وما وراء المحركات، بوابات الويب. وسوف نتناول هذه الأدوات بشكل أكثر تفصيلاً في فصل مستقل للتعرف إلى طريقة بناء هذه الأدوات وآليات عملها في التكشيف والتحليل والبحث والفرز.

◄ 3.4.2 التقسيم وفقاً لأنواع المداخل المكشفة

تتنوع مداخل التكشيف بتنوع الوحدات المكشفة، والتي تحدد المدخل الملائم لترتيب التسجيلات التي تتضمنها الكشافات. وعلى الرغم من أن قضية الترتيب لم تعد بالأهمية التي كانت عليها قبل استخدام أنظمة التكشيف الآلية التي أصبحت

الأساس الآن في إعداد الكشافات، إلا أن بنية هذه النوعية من الكشافات كان لها أثر كبير في تطور أساليب التكشيف وبنية الكشافات الآلية وقواعد البيانات. ويمكن تقسيم الكشافات وفقاً لنوعية مدخل التكشيف إلى:

3.4.2.1 كشافات العناوين

هي الكشافات التي تركز على عناوين الأعمال من كتب ومقالات وأعمال مؤتمرات. وقد ظهرت أول أشكال كشافات العناوين مع بداية استخدام نظام المصطلح الواحد Uni-Term في إعداد كشافات التباديل الموضوعية للعناوين. فظهرت أنواع عدة من الكشافات التي تركز على استخدام المصطلحات الواردة في العناوين للدلالة على المحتوى الموضوعي للوثائق. ويعد كشاف الكلمات المفتاحية في السباق Key Words In Context (KWIC) أبرز مثال لهذه النوعية من الكشافات. بالتالي فإن مصطلحات عناوين الوثائق تستخدم كمؤشر للدلالة على المحتوى الموضوعي للوثائق. كما تستخدم كمداخل لترتيب هذه النوعية من الكشافات.

3.4.2.2 كشافات الموضوعات

تعد هذه الفئة أشهر أنواع الكشافات وأكثرها انتشاراً واستخداماً، حيث إن قواعد البيانات الببليوجرافية المتخصصة في المجالات الموضوعية المختلفة ما هي إلا كشافات موضوعية متاحة في شكل إلكتروني. وتوجد نماذج عدة من هذه الكشافات متاحة في صورة قواعد بيانات ببليوجرافية وقواعد بيانات للنصوص الكاملة التي تصدر عن الناشرين التالين:

Elsevier - https://www.elsevier.com

Springer – https://www.springer.com

Wiley – https://www.wiley.com

وغيرهم مثل: OVID, TAYLOR and Francis, EMARLD, SAGE،.. إلخ.

وعلى المستوى العربي بدأت الكثير من الشركات العربية مع بداية الألفية الجديدة

في إنشاء قواعد بيانات بالمحتوى العربي في مختلف التخصصات. نذكر منها على سبيل المثال لا الحصر:

1. دار المنظومة: http://www.mandumah.com
2. المنهل: https://www.almanhal.com
3. مكتبة دبي الرقمية https://ddl.ae
4. إثراء المعارف الرقمية http://ethraadl.com
5. معرفة http://www.e-marefa.net/ar

3.4.2.3 كشافات المؤلفين

تعد قوائم الأسماء والأعلام الواردة في الأعمال العلمية والأدبية من الأدوات المهمة التي يحتاج الأفراد والمؤسسات إلى إبرازها. لذلك اهتمت العديد من المؤسسات بإعداد كشافات بالمؤلفين والتي كانت تظهر في نهايات الكتب أو المواد المرجعية مثل الموسوعات، وترتب ترتيباً هجائياً وفقاً لأسماء المؤلفين المستشهد بأعمالهم الأدبية والعلمية في متن النص. ومع تطور منصات البحث المتاحة على الخط المباشر، أتاحت تلك المنصات إمكانيات البحث بأسماء المؤلفين للوصول إلى كافة أعمال مؤلف معين، كما هو الحال في قواعد بيانات الاستشهادات المرجعية التي سبق ذكرها. وبظهور وتطور تلك المنصات اختفت تقريباً كشافات المؤلفين المستقلة وأصبح الاعتماد بصورة أكبر على تلك المنصات في التعرف إلى أعمال المؤلفين وتقييم أدائهم العلمي والمعرفي. كما ظهرت أدوات جديدة في البيئة الرقمية تتسم بملامح المشابكة والتواصل بين المؤلفين والباحثين، عرفت بشبكات المؤلفين الاجتماعية والتي تم تطبيقها في القياسات البديلة كما سنوضح لاحقاً. وقد بدأت العديد من المؤسسات البحثية والأكاديمية أخيراً، الاهتمام بإعداد ملفات السمات الأكاديمية E-protoflio ليوفر بيانات كاملة عن كافة الباحثين المنتمين لتلك المؤسسات.

ونظراً لأهمية دور المؤلفين ومشاركتهم العلمية والحاجة إلى تقييم أدائهم، ظهرت العديد من مؤشرات القياس التي تحاول وضع مقاييس رقمية لتقييم الإنتاجية العلمية للمؤلفين وأثرهم في المجالات البحثية. وتنقسم هذه القياسات إلى نوعين رئيسين:

أ. مقاييس ببليومترية

وتعتمد تلك المقاييس على مؤشرات الإنتاجية العلمية وجودة الإنتاج العلمي الذي يتم قياسه من خلال معدلات الاستشهاد. وقد تم ابتكار العديد من المؤشرات لقياس الأداء العلمي للمؤلفين لعل أبرزها:

- **كشاف H Index**: وهو مقياس ابتكره العالم هيرش Hirsh ليحدد درجة مساهمة المؤلف بناء على عدد المقالات المنشورة وعدد الاستشهادات التي حصلت عليها. ووفقاً لهذا الكشاف يحصل المؤلف على كشاف h إذا كان حصل عدد من أبحاثه على حد أدنى يعادل ترتيب البحث في القائمة التنازلية. فعلى سبيل المثال يحصل الباحث على كشاف h يعادل 5 إذا حصل 5 أبحاث من قائمة أبحاثه على 5 استشهادات على الأقل. ولإجراء عملية القياس بدقة يتم ترتيب قائمة الأبحاث ترتيباً تنازلياً، وفقاً لعدد الاستشهادات. وتكون قيمة h تعادل قيمة الأبحاث N التي حصلت على N من الاستشهادات أو أكثر.

- **مقياس I 10 Index**: وهو مقياس يطبقه جوجل العلمي منذ عام 2011 لتحديد عدد الأبحاث التي حصلت على الأقل على عدد 10 استشهادات كمقياس لجدارة الأعمال، حيث اعتبرت أن حصول البحث على عدد 10 استشهادات مقياس جدارة، أما الأبحاث التي تحصل على عدد أقل من 10 استشهادات لا تدخل في قائمة التقييم. من ثم فمؤشر I 10 Index يعتمد على إحصاء عدد المقالات التي نشرها الباحث خلال فترة زمنية ثم إحصاء عدد المقالات التي تمثل المؤشر I التي حصلت على 10 استشهادات على الأقل. ولعل أبرز مزايا هذا المقياس أنه بسيط في طريقة حسابه ويضع مؤشراً للجدارة يراعي كفاءة كل بحث على حدة. ويمكن من خلاله تقييم أداء الباحثين خلال فترة

زمنية، إلا أنـه يفتقـر إلـى وجـود دلالـة واضحـة لشـكل مخرجـات المؤلـف بصفـة عامـة Author Contribution Shape.

وعلى غـرار هذيـن المقياسيـن تـم ابتكـار عـدد آخـر مـن المقاييـس التـي تحـاول التغلب علـى بعـض الصعوبـات التـي توجـد فـي المقياسيـن السـابقين ومنهـا: ,G index, A index h5 index, P 100. وتعتمـد كل هـذه المقاييـس علـى نمـط القيـاس نفسـه المطبـق فـي كشـاف H وفـي مقيـاس I 10، مـن حيـث تطبيـق مؤشـر لتحديـد حجـم المسـاهمة العلميـة بنـاء علـى مقيـاس للجـدارة والاسـتحقاق الأكاديمـي.

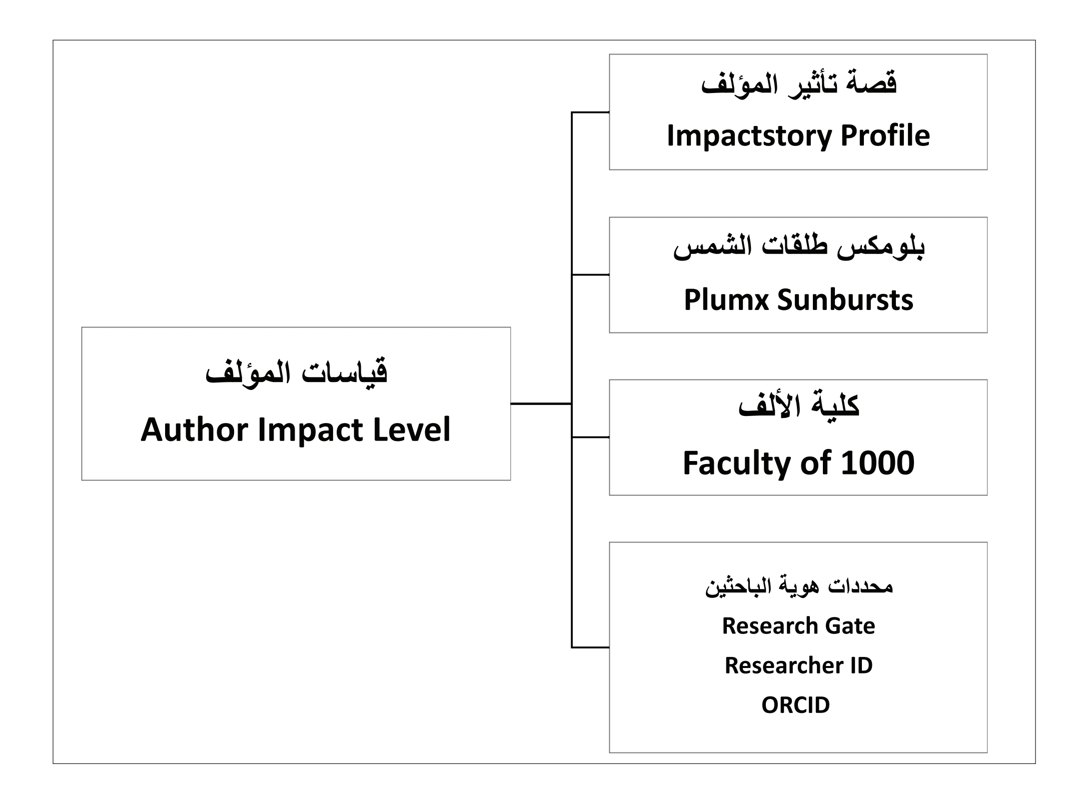
II. مقاييس بديلة

ظهـرت فكـرة المقاييـس البديلـة علـى يـد جاسـون بريـم Jason Priem فـي عـام 2010 الـذي كان طالـب دراسـات عليـا بجامعـة نـورث كارولينـا بتشـبل هيـل، والـذي نشـر بحثـاً بعنـوان Altmetrics: A Manifesto. تسـتند هـذه النوعيـة مـن القياسـات إلـى تحليـل الويـب الاجتماعـي Social Web. يشـتمل هـذا المقيـاس علـى ثلاثـة ملامـح:

- العمل في بيئة الويب
- الحاجة الماسة إلى قياسات جديدة وتوافر بيانات مهمة تدعم هذه القياسات
- القياسات البديلة مرتبطة بأنشطة الاتصال العلمي

وتعـد القياسـات البديلـة امتـداداً لحركـة التجميـع والمتابعـة والتحليـل للأنشـطة العلميـة بغـرض التقييـم والترتيـب، ولا تقتصـر علـى المـواد التقليديـة مثـل الكتـب والدوريـات، ولكـن تشـمل أيضـاً العـروض والملصقـات والمحاضـرات المسـجلة والتعليقـات والمدونـات والتدويـن الصوتـي Podcast الفديوهـات والرسـوم البيانيـة ومجموعـات البيانـات Datasets.

وقـد أتاحـت أدوات الجيـل الثانـي للويـب أسـاليب أكثـر مرونـة وسـرعة للحـوار والنقـاش داخـل وخـارج المجتمـع العلمـي. ويوجـد أربعـة قياسـات بديلـة للمؤلفيـن يوضحهـا الشـكل التالـي:



Research Gate: https://www.researchgate.net

Researcher ID: https://clarivate.com/products/researcherid

ORCID: https://orcid.org

3.4.2.4 كشافات الكيانات

هي قوائم بأسماء الهيئات أو الأماكن أو المؤسسات أو العناصر الكيميائية والعلامات التجارية وغيرها من الكيانات التي ترد في متن الأعمال. ويهتم العديد من المؤلفين بإعداد كشافات بالمختصرات والاستهلاليات المستخدمة للدلالة على أسماء الكيانات الواردة في أعمالهم. كما توجد مجموعة أخرى من الكشافات ولكنها أقل انتشاراً، من المجموعة السابق ذكرها، مثل كشافات المعادلات والتركيبات (الكيميائية والرياضية) كشافات التواريخ والأرقام، كشافات الأجناس والفئات.. وغيرها.

3.4.3 تقسيم الكشافات وفقاً لطريقة الترتيب

توجد ثلاث طرق أساسية لترتيب المواد في الكشافات وغيرها من أدوات التمثيل والضبط المتاحة في شكل مطبوع أو رقمي، هي: الترتيب الهجائي، الترتيب المصنف، الترتيب القاموسي.

3.4.3.1 الترتيب الهجائي

توجد طريقتان أساسيتان بصفة عامة للترتيب الهجائي، الأولى تعتمد على الترتيب كلمة بكلمة Word By Word، وفي هذه الحالة فإن كلمة مثل San Salvador سوف تسبق كلمة مثل Sandman على أساس أن San كلمة منتهية. أما الطريقة الثانية فتعتمد على الترتيب حرف بحرف Letter By Letter وفي هذه الحالة فإن Sandman سوف تسبق San Salvador على اعتبار أن حرف d يسبق في الترتيب الحروف الخاصة مثل المسافات وغيرها. كما أن كلمة مثل «استراتيجية» سوف تسبق «استراتيجيات سياسية» في نظام ترتيب كلمة بكلمة بينما تسبق «استراتيجيات سياسية» كلمة «استراتيجية» في نظام ترتيب حرف بحرف.

3.4.3.2 الترتيب المصنف

يعتمد الترتيب المصنف على تطبيق نظام التقسيم إلى فئات من خلال تطبيق خطط تصنيف المعرفة ومنها خطط التصنيف العامة مثل تصنيف ديوي العشري، العشري العالمي، مكتبة الكونغرس؛ أو تطبيق نظام تصنيف متخصص. توجد طريقتان أساسيتان لإعداد الكشافات المصنفة، في الطريقة الأولى تظهر المداخل تحت أرقام مخصصة ودقيقة إلى حد كبير، وتشتق هذه الأرقام من خطة تصنيف عامة أو متخصصة. وهذه الطريقة كانت الطريقة الأساسية في إعداد وتجهيز المداخل الموضوعية، حيث ترتب المداخل الموضوعية وفقاً لخطة تصنيف وجهية Faceted Classification Scheme معدة خصيصاً للتطبيق في الكشاف. كما توجد بعض الكشافات المطبوعة التي تعتمد على نظم تصنيف عامة مثل خطة تصنيف العشري العالمي Universal Decimal Classification (UDC).

أما الطريقة الثانية لبناء الكشافات المصنفة فتستخدم في ترتيب المداخل الموضوعية بالكشافات، وتعتمد على اشتقاق الرؤوس الموضوعية من قواعد البيانات، ثم يتم تجميع المداخل تحت فئات موضوعات عريضة مرتبطة، بالتالي يمكن الوصول إلى رؤوس الموضوعات الدقيقة من خلال الكشافات المساعدة، حيث ترتب المداخل تحت فئات موضوعية عريضة وتحت كل فئة موضوعية توجد فئات ثانوية. وقد استخدم هذا النمط من الترتيب أيضاً في بناء أدلة البحث لمصادر الويب التي سوف نناقشها بالتفصيل فيما يلي، وعادة ما يكون ناتج عملية التكشيف والترتيب في حالة الاعتماد على الترتيب المصنف أحد أنواع الكشافات المعروف بالكشاف المتسلسل Chain Index.

- **الكشاف المتسلسل Chain Indexing**

يستخدم هذا النمط من أنماط التكشيف لمعالجة وترتيب رؤوس الموضوعات التي يتم اشتقاقها من خلال خطط التصنيف عامة أو متخصصة. والهدف من إعداد هذا النوع من الكشافات ضمان توافر مداخل تحت كل مصطلح من المصطلحات المكونة للرأس المركب، فضلاً عن ربط هذه المداخل في سلسلة بالمصطلحات الأعرض والأضيق منه في البناء الهرمي. معنى ذلك أن المصطلحات في الكشاف المتسلسل تظهر في شكل سلسلة تنتقل من العام إلى الخاص.

3.4.3.3 الترتيب القاموسي

يشير هذا النوع من أنواع الترتيب إلى نمط الترتيب المستخدم في القواميس الهجائية، ويتميز باستخدام كل المداخل بكافة أنواعها من مؤلفين وعناوين وموضوعات في ترتيب هجائي واحد. ويتنوع الترتيب في هذه الحالة أيضاً ما بين الترتيب كلمة بكلمة أو الترتيب حرفاً بحرف. وعادة ما يستخدم الترتيب القاموسي في إعداد الكشافات التجميعية Cumulative Indexes التي تتضمن مداخل المؤلفين والهيئات والمؤسسات في كشاف واحد. ويمكن إعداد هذا النوع من الكشافات للكتب والمصادر المرجعية مثل الموسوعات وأدلة العمل والكتب السنوية

وصفحات الويب الصفراء Yelow Web Pages. كما تم استخدام هذه الطريقة في إعداد الفهارس القاموسية قبل ظهور الفهارس المتاحة على الخط المباشر. وهي نوع متميز من الفهارس اليدوية كانت ترتب فيه كل أشكال المداخل في ترتيب هجائي واحد، مع إعداد الإحالات المناسبة وخاصة إحالة (انظر أيضاً)، حيث إنه يمكن أن يكون لكل عمل على الأقل مدخل رئيس بالمؤلف وآخر بالعنوان وثالث بالموضوعات. بالتالي يتم إعداد حالات (انظر أيضاً) إلى مواقع البطاقات الخاصة والموضوعات في ترتيب بطاقة المؤلف. ويساعد هذا النوع من الترتيب على سهولة الوصول إلى مصادر المعلومات، إلا أنه يعيبه كبر حجمه وصعوبة إعداده. ومع ظهور أدوات البحث في البيئة الرقمية اختفت هذه النوعية من أساليب الترتيب اليدوي وظل مفهوم الترتيب مستخدماً في البيئة الرقمية في الأدلة والوكيبيديا والموسوعات الرقمية والكشافات التجميعية.

3.5 قضية التمثيل

تم استعراض الطرق المختلفة لتمثيل المعلومات وتصنيفاتها المتنوعة، والتي تشمل التكشيف الاستخلاص والملخصات والاشتقاقات والتقسيم إلى فئات والتوسيم الاجتماعي والملخص الوافي للموقع. وتعد هذه الأساليب أبرز الطرق المعروفة لتمثيل المعلومات في البيئة الرقمية، كما تم توضيحه مسبقاً فإن هذه الطرق تختلف عن بعضها بعضاً في مدى تمثيلها للوثيقة الأصلية. وعند ترتيب الفئات الخمس من حيث جودة التمثيل يأتي التكشيف على قمة هذه الفئات يليه الاستخلاص من حيث الأهمية والتطبيق أيضاً، ويعد موم أقل هذه الفئات استخداماً ثم يأتي كل من التقسيم إلى فئات والتوسيم الاجتماعي في منطقة متوسطة بينهما. مع العلم أن التوسيم الاجتماعي بدأ يزداد الاهتمام به في السنوات الأخيرة مع زيادة الاهتمام بتطوير الويب الدلالي وأدوات التفاعل الاجتماعي. وعلى الرغم من أن هذه المقارنة موجزة، إلا أنها تلقي الضوء على كيفية استخدام كل طريقة من طرق تمثيل المعلومات لأداء مهمة تيسير سبل الوصول إلى المعلومات.

3.6 الطرق الأخرى لتمثيل المعلومات

تعد عمليات التكشيف والتقسيم إلى فئات والتلخيص، أساليب تقليدية لتمثيل المعلومات؛ وإلى جانب هذه الطرق التقليدية توجد مجموعة من الأساليب الفريدة في نوعها من حيث آليات تمثيلها للمعلومات وفي طريقة تطبيقها واستخداماتها في تمثيل المعلومات والتي سيتم مناقشتها في هذا الجزء.

3.6.1 الاستشهادات Citations

تشير الاستشهادات إلى المصادر التي يرجع إليها المؤلف عند إعداده لبحث أو دراسة، ويستعين بها في كتابته العلمية. وقد عرفت في تاريخ العلوم بعلم السند Authenticity الذي يهتم بتوثيق المعلومات ومصادرها وجودة تلك المصادر. والاستشهاد يعني بصفة عامة توثيق العلاقة بين كل أو جزء من الوثيقة المُسْتَشْهَدِ بها Cited Document وكل أو جزء من الوثيقة المُسْتَشْهِدَة (Malin,1968) Citing Document. فمنذ أن ابتكر Dr. Eugene Garfield فكرة الاستشهادات وطرق قياسها في العصر الحديث، وأسس معهد المعلومات العلمية، قام بنشر كشافات الاستشهادات المرجعية والتي تشمل:

- كشاف استشهادات العلوم Science citation Index.
- كشاف استشهادات العلوم الاجتماعية Social science citation Index.
- كشاف الإنسانيات والفنون Arts & Humanities citation Index.

وقد كان لظهور كشافات الاستشهادات المرجعية أثر كبير في تطوير أدوات قياس القيمة العلمية لمصادر المعلومات وتمثيلها بأرقام تدل على أهميتها العلمية من خلال معدلات الاستشهادات المرجعية بتلك المصادر. ولعل أهم هذه الأدوات تقرير الاستشهادات المرجعية Journal Citation Report والذي يقوم بترتيب الدوريات العلمية وفقاً لأهميتها النسبية وقيمتها المعرفية من خلال عدد مرات الاستشهاد بها. كما ظهرت في السنوات الأخيرة كشافات لاستشهادات المؤتمرات العلمية في مجالات العلوم والعلوم الاجتماعية والإنسانيات.

Scientific Conference Proceedings Citation Index

Social Science and Humanties Citation Index

ومنذ بداية نشر كشافات الاستشهادات المرجعية في منتصف الستينات من القرن الماضي إلى الآن ويوجد جدل دائر حول أهمية الاستشهادات المرجعية ومدى مصداقيتها كأداة لقياس القيمة المعرفية لمصادر المعلومات، ويمكن تلخيص تلك الأسباب في تكريم الرواد ومنحهم حقوقهم الأدبية في الأعمال المنسوبة إليهم إلى جانب التعرف إلى القيمة العلمية والمعرفية للوثائق والمصادر والمؤسسات. ومنذ ظهور كشافات الاستشهادات المرجعية واستخدامها بدأت العديد من قواعد البيانات تهتم برصد الاستشهادات في صورة إلكترونية وإعداد إحصاءات دقيقة بمعدلات الاستشهاد العلمي لعل أهمها:

- **شبكة المعرفة بمعهد المعلومات العلمية ISI Web of Knowledge**

تعد شبكة المعرفة إحدى أهم وأقدم قواعد بيانات التكشيف والاستشهادات المرجعية في العالم، حيث نشرت لأول مرة في صورة مطبوعة في عام 1964 تحت مسمى كشاف استشهادات العلوم Science Citation Index وقد ابتكرها الدكتور يوجين جارفيلد الذي أسس فيما بعد المعهد القومي للمعلومات Institute of Scientific Information – ISI، كما سبق وذكرنا، لكي يقوم بحصر وتكشيف وإنتاج كشافات الاستشهادات المرجعية فيما بعد. وقد تم بيع عنكبوت المعرفة إلى مجموعة شركات رويترز، فظهرت تحت اسم مؤسسة تومسون رويترز Thomson Reuters والتي تتولى إصدار مجموعة مهمة من المنتجات التي تساعد على تتبع حركة النشر الدولي بصورة دقيقة. ومن أهم مخرجات هذه المؤسسة شبكة العلوم ISI Web of Science والتي تشتمل على عدد كبير من المخرجات العلمية الدولية التي يتم تكشيفها وتحليلها للتعرف إلى توجهات النشر الدولي في مختلف مجالات العلوم والمقارنة. وتشتمل قاعدة بيانات شبكة العلوم على المواد التالية:

- 23 ألف دورية علمية
- نحو 23 ألف براءة اختراع

- 110 آلاف أعمال مؤتمرات
- 9 آلاف موقع ويب
- أكثر من 40 مليون تسجيلة لتلك المواد مجتمعة
- يمكن بحث كافة تلك المصادر بصورة كاملة من خلال صندوق بحث واحد.

- **المستكشف Scopus**

ظهرت قاعدة بيانات SCOPUS كمنافس لقاعدة بيانات عنكبوت العلوم منذ عام 1997 وبدأت في تكشيف أكثر من 25 ألف دورية علمية. وهو ثاني أكبر قواعد بيانات الاستشهادات المرجعية التي يمكن من خلالها التعرف إلى توجهات النشر الدولي وتأثير الدول في المجالات العلمية المخلتفة. وتتميز تلك القاعدة بتركيزها بشكل عميق على تقييم الباحثين وإعطاء بطاقة هوية كاملة لكل باحث، تحدد معدلات النشر التي قام بها وعدد المصادر التي اعتمد عليها وعدد الاستشهادات التي حصل عليها وتاريخه المهني والأكاديمي، ما يجعلها أداة مهمة لتقييم الباحثين على المستويات المحلية والإقليمية والدولية.

تتضمن مستخلصات واستشهادات مرجعية حول الإنتاج الفكري المنشور في الدوريات العلمية ومصادر الويب في جميع مجالات المعرفة البشرية. كما تساعد على التعرف إلى الإنتاج الفكري المنشور في أكثر من 15 ألف عنوان متاح لدى أكثر من 4000 ناشر، كما تشتمل على أكثر من 12850 دورية أكاديمية، 500 دورية منشورة على الويب، ملخصات واستشهادات 700 مؤتمر علمي، 28 مليون مستخلص، 245 مليون استشهاد مرجعي، 13 مليون براءة اختراع.. الخ. /http://www.scopus.com scopus/home.url

وقد أدت المنافسة بين المصدرين السابقين (عنكبوت المعرفة والمستكشف) إلى سباق في تقديم مؤشرات وأساليب عرض جديدة لتقييم الأداء العلمي والقيمة البحثية لمصادر المعلومات لعل أهمها:

- **معامل التأثير Impact Factor** والذي يعد الأداة الأساسية في تقييم الدوريات العلمية وجودة وكفاءة النشر العلمي في مختلف آليات القياس العالمية.

- **كشاف H H Index** الذي أصبح يستخدم لكل من الأفراد والدوريات في عمليات التقييم وكفاءة آليات القياس.

- **تطبيع تأثير المصدر لكل وثيقة Source Mormalized Impact Per Paper-SNIP**. وهو عبارة عن مقياس لمعدلات الاستشهاد بحسب عدد مرات الاستشهاد بكل مقالة مع الأخذ في الاعتبار النوع في معدلات الاستشهاد من مجال إلى آخر.

وعلى الرغم من الاختلافات بين التخصصات من حيث فرص الاستشهاد، والمؤشرات التي يتم على أساسها تحديد الأهمية العلمية وفقاً لعدد الاستشهادات، إلا أن الدافع وراء الاستشهاد بأعمال الآخرين قد يختلف من باحث لآخر. ويمكن النظر إلى الاستشهاد على أنه اختيار من جانب الباحث لمجموعة من الوثائق تمثل بحثه، وعملية التمثيل تأخذ في هذه الحالة شكل الاستشهادات بدلاً من بدائل الوثائق التقليدية مثل المستخلصات، ومصطلحات الكشاف. فالاستشهادات عبارة عن بيانات ببليوجرافية مثل المؤلف أو المؤلفين والعنوان وبيانات.. الخ. وتعبر عن وثائق تم الاستشهاد بها، بمعنى أنه لا توجد حاجة إلى بناء وصيانة أدوات أخرى مثل المكانز، وخطط التصنيف لأغراض تمثيل المعلومات، حيث يكتفي بالبيانات الببليوجرافية لكي تعبر عن الوثيقة.

وتعتمد عملية الاستشهاد على قيام المؤلف باختيار مجموعة من الوثائق يستشهد بها لكي تعبر عن وثيقة من خلال قائمة المصادر References، من ثم فهو يقوم بعملية التمثيل بنفسه. وقيام المؤلف بهذه العملية يعني التخلص من دور الوسيط في عملية التمثيل، ما يكون له تأثيرات إجابية وأخرى سلبية، لعل أبرز التأثيرات الإيجابية أن المؤلف هنا يقوم بدور الكشف وهو على دراية أكبر بالوثيقة وليس بحاجة إلى بذل جهود إضافية لتفسير الوثيقة الأصلية، أما التأثير السلبي فيرجع إلى أنه لا يوجد تفسير واضح لأسباب الاستشهاد بوثيقة ما وعدم الاستشهاد بأخرى.

ومـن الأمـور التـي تثيـر الكثيـر مـن التسـاؤلات حـول الاستشـهادات كأداة لتمثيـل المعلومـات هـو مـدى التغطيـة وحـدود التغطيـة لقواعـد بيانـات الاستشـهادات. ومـع ذلـك فـإن الباحثيـن فـي حاجـة ماسـة إلـى اسـتخدام تلـك المصـادر، نظـراً لأن بنـاء قاعـدة بيانـات جديـدة أمـر فـي غايـة الصعوبـة ويسـتغرق وقتـاً طويـلاً. كمـا أن عمليـة تكشـيف الاستشـهادات المرجعيـة لا تتطلـب أي معرفـة خاصـة أو ذكاء بشـري؛ لذلـك فإنـه مـن الممكـن ميكنـة العمليـة بالكامـل ودون تدخـل مـن جانـب البشـر، والـذي يبـدو أنـه لا يمكـن تحقيقـه مـع الأسـاليب الأخـرى لتمثيـل المعلومـات.

3.6.2 تكشيف سلاسل الحروف
Strings Indexing

السلاسـل عبـارة عـن مجموعـة مـن الجمـل والعبـارات التـي يتـم تكشـيفها لتمثيـل وثيقة مـا. وتوجـد أنمـاط متعـددة لتكشـيف السلاسـل تجمعهـا كلهـا خاصيتـان أساسـيتان همـا:

1. تتم عملية التكشيف بصورة يدوية لتحديد سلسلة الحروف التي تمثل وثيقة ما.
2. يتـم تجميـع مداخـل الكشـاف بطريقـة آليـة بالإعتمـاد علـى سلسـلة الحـروف التـي تـم إعدادهـا لتمثيـل الوثيقـة.

لذلـك، يمكـن اعتبـار تكشـيف الحـروف أحـد أنمـاط الكشـافات الآليـة التـي تـم وصفهـا سـابقاً. وتُعـد كشـافات الكلمـات المفتاحيـة Key Words In Context أحـد أبـرز نمـاذج كشـافات السلاسـل ومثـال لهـا الكشـاف المعـروف بنظـام كشـاف السـياق المحفـوظ Preserved context index system - PRECIS ونظـام تكشـيف العبـارات المتضمنـة (Nested Pharse Index system (NEPHIS. وفـي هذيـنِ النظاميـن يقـوم المكشـف يدوياً بتحديـد سلسـلة حـروف فـي صـورة عبـارة أو جملـة للتعبيـر عـن الوثيقـة، ثـم يتـم تكشـيفها كلمـة بكلمـة مـن خـلال النظـم الآليـة. ففـي نظـام PRECIS يتـم إعـداد شـبه مسـتخلص يتـم تكشـيفه باسـتخدام الكلمـات المفتاحيـة الـواردة فيـه، ويعتمـد نظـام NEPHIS علـى اسـتخدام الملخـص أو موجـز يختـاره المكشـف مـن الوثيقـة للدلالـة عليهـا، ثـم يتـم تكويـد هـذه السلاسـل لتحديـد المصطلحـات التـي تصلـح أن تُسـتخدم كلمـاتٍ مفتاحيـةً لكـي يتـم

توظيفها كمداخل بالكشافات. وبناء على ذلك فإن الجزء الآلي في عملية تكشيف السلاسل يمكن أن يتم معالجته آلياً بسهولة وكفاءة كبيرة.

ويساعد التكامل بين التدخل البشري في اختيار العبارات والجمل الممثلة للوثائق مع استخدام النظم الآلية في أداء الجزء الميكانيكي في العملية، على جعل تلك العملية تحمل الكثير من المزايا والجاذبية في تمثيل الوثائق. فهي من ناحية تحافظ على جودة عملية التكشيف نظراً للتدخل البشري في الاختيار الدقيق للعبارات والجمل التي تمثل الوثائق، ومن ناحية أخرى، فهي تمنع أو تتخلص من كل الإجراءات المملة وغير الفعالة، والتي لا تساعد على تحقيق الاطراد في التكشيف بالنظم اليدوية من خلال الاعتماد على آلية موحدة بالنظم الآلية. لذلك فإن التطور السريع في المعلومات الرقمية سوف يؤدي إلى انتشار استخدام النظم الآلية في تمثيل المعلومات وفي استرجاعها أيضاً.

3.7 ملخص للاتجاهات الأساسية في تمثيل المعلومات

اشتمل هذا الفصل على شرح مفصل للطرق والأساليب المختلفة لتمثيل المعلومات ويوضح الجدول 2.1 الاتجاهات الأساسية التي تمت مناقشتها في هذا الفصل، سواء من حيث نوع التمثيل (استخدام لغة مضبوطة أو حرة في التكشيف) إلى جانب طريقة الإنتاج وكل طريقة من هذه الطرق لها مزاياها وعيوبها. ويشير إلى أنه عند اختيار طريقة معينة لتمثيل المعرفة فسوف تقوم بتحقيق ما يلي:-

1. التمييز بين المداخل المختلفة.
2. تحديد المداخل المتشابهة.
3. إعداد وصف دقيق للمداخل.
4. إزالة أو تحليل حجم الغموض عند التفسير.

وبالطبع لا يمكن لطريقة واحدة أن تحقق كل المتطلبات اللازمة لعملية التمثيل، حيث إن إحدى نقاط الضعف في طريقة ما، قد تكون ميزة كبرى في طريقة

أخرى. لذلك فالتكشيف وحده مثلاً من الممكن أن يوضح الموضوعات المحددة التي تعالجها الوثيقة، إلا أن المستخلص يوضح مضمون الوثيقة ككل. لذلك فإن التعددية في الأساليب والطرق Methodological Pluralism تُعد أفضل الوسائل لتمثيل المعلومات بدقة وكفاءة. فالمزج بين طرق التمثيل المختلفة مثل التصنيف والاستخلاص والتكشيف والتوسيم يمكن أن يحقق العديد من المزايا التي تفوق استخدام طريقة واحدة.

ويوجد تطور سريع في استخدام الأساليب الحديثة المصاحبة للجيل الثاني للويب الذي يعتمد على مشاركة المستفيد في عمليات التطوير والبناء مثل التلخيص الوافي للمحتوى أو التوسيم RSS &Tagging وذلك بغرض تحقيق الاحتياجات الجديدة لتمثيل المعلومات في العصر الرقمي.

المصادر

- الهجرسي، سعد محمد (1980). الإطار العام للمكتبات والمعلومات: أو نظرية الذاكرة الخارجية. القاهرة: جامعة القاهرة، 58 ص.
- الهجرسي، محمد سعد (1991). المكتبات والمعلومات: أسس علمية حديثة ومدخل منهجي عربي. الرياض: دار المريخ.
- لانكستر، ولفرد (1997) أساسيات استرجاع المعلومات/ ترجمة حشمت قاسم، الرياض: مكتبة الملك فهد الوطنية، 454 ص.
- عبد الهادي، محمد فتحي (2005). التكشيف والاستخلاص. القاهرة، الدار المصرية اللبنانية، 284 ص.
- حسام الدين، مصطفى. مجموعة محاضرات في استرجاع المعلومات، جامعة القاهرة 1994.

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 34-43.
- Chu, H. (2001). Research in image indexing and retrieval as reflected in the literature. Journal of the American Society for Information Science and Technology, 52(12), 1011-1018.
- Chu, H., & Rosenthal, M. (1996, October). Search engines for the World Wide Web:

A comparative study and evaluation methodology. In Proceedings of the Annual Meeting-American Society for Information Science (Vol. 33, pp. 127-135).Dempsey, L., & Heery, R. (1998). Metadata: a current view of practice and issues. Journal of documentation, 54(2), 145-172.

- Fugmann, R. (1993). Subject analysis and indexing: theoretical foundation and practical advice (Vol. 1). Indeks Verlag Dr. Ingetraut Dahlberg.
- Graf, Peter, and D. Fehrer. "Term indexing." Automated Deduction—A Basis for Applications. Springer, Dordrecht, 1998. 125-147.
- - Hardin (Ed.), Proceedings of the 59th Annual Meeting of the American Society for Information Science (pp. 127-135). Medford, NJ: American Society for Information Science.
- International DOI Foundation. (2009). Welcome to the DOI system. Retrieved july 18.2009. from www.doi.org
- Jones, K. Sparck. "Automatic summarizing: factors and directions." Advances in automatic text summarization (1999): 1-12.
- Jones, K. S., Jones, G. J. F., Foote, J. T., & Young, S. J. (1996). Experiments in spoken document retrieval. Information Processing & Management, 32(4), 399-417.
- Kelly, Brian. (2005). RSS: Moe than just news feeds. New review of information Network,11(2), 219 227.
- Knight, K. (1999). Mining online text. Communications of the ACM, 42(11), 58-61.
- Lancaster, Frederick Wilfrid, et al. Indexing and abstracting in theory and practice. London: Library Association, 1991.
- Lassila, O. (1997). Introduction to RDF Metadata–W3C Note. World Wide Web Consortium, Cambridge, MA. URL http://www. w3. org/TR/NOTE-rdf-simple-intro-971113. html.
- Lesk, Michael. (1997). Practical digital libraries: Books, Bytes and bucks. San Francisco: Morgan Kaufmann.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.
- Malin, M. V. (1968). Science Citation Index: a New concept in indexing. Library Trends, 16(3), 374-374.
- Mathes, A. (2004). Folksonomies-cooperative classification and communication through shared metadata: December.
- Meadow, C. T., Boyce, B. R., & Kraft, D. H. (1992). Text information retrieval systems (Vol. 20). San Diego, CA: Academic Press.
- Nair, S. S., & Jeeven, V. K. J. (2003). A brief overview of metadata formats. DESIDOC Journal of Library & Information Technology, 24(4).
- Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. Journal of information science, 20(2), 108-118.
- Rowley, J., & Hartley, R. (2017). Organizing knowledge: an introduction to managing access to information. Routledge.

- Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. IEEE intelligent systems, 21(3), 96-101.
- Shafer, K. E. (2001). Mantis project: A toolkit for cataloging. Journal of library administration, 34(3-4), 339-344.
- Vizine Goetz, D. (1997). From book classification to knowledge organization: improving Internet resource description and discovery. Bulletin of the American Society for Information Science and Technology, 24(1), 24-27.
- Wang, J. (2007). Digital object identifiers and their use in libraries. Serials review, 33(3), 161-164.
- Weibel, S. (1997). The Dublin Core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science and Technology, 24(1), 9-11.
- Wool, G. (1998). A meditation on metadata. The Serials Librarian, 33(1-2), 167-178.
- Zeng, lei; Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. Journal of the American Society for information science and technology, 55(5), 377-395.
- Zeng, M. L. (2008). Metadata. Neal-Schuman Publishers, Inc..
- Zhang, H., Low, C. Y., Smoliar, S. W., & Wu, J. (1995, January). Video parsing, retrieval and browsing: an integrated and content-based solution. In Proceedings of the third ACM international conference on Multimedia (pp. 15-24). ACM.

الفصل الرابع

مصادر البيانات
بنظم تمثيل المعرفة

◄ 4 مقدمة

يتناول هذا الفصل مصادر البيانات المرتبطة بعمليات تمثيل المعلومات والمعرفة، حيث سيتم مناقشة أنواع البيانات وفئاتها والميتاداتا وطرق تمثيلها والنصوص الكاملة، والبيانات المستخدمة في تمثيل الوسائط المتعددة.

◄ 4.1 أنواع البيانات

يتم تقسيم البيانات إلى ثلاثة أنواع أساسية هي كالتالي: غير مهيكلة Unstructured، شبه مهيكلة semistructured، مهيكلة structured (محمد وآخرون، 2018). ولكل نوع من تلك الأنواع الثلاثة إطار تحدده الوظائف التي يسعى لتحقيقها. تظهر البيانات غير المهيكلة في صورة غير نمطية ليس لها شكل أو حجم محدد، حيث إنها كيانات ليس لها إطار ثابت يجمعها أو شكل موحد. وعلى الطرف الآخر، تظهر البيانات المهيكلة في صورة نمطية من خلال أطر محددة، فهي عبارة عن بيانات لها نمط ثابت بحيث يمكن تخزينها في قاعدة بيانات وكل عنصر بيانات منها له شكل وإطار نمطي مميز. وسيتم فيما يلي مناقشة الأنواع الثلاثة للبيانات.

◄ 4.1.1 البيانات غير المهيكلة
Unstructured Data

تتميز هذه النوعية من البيانات بأنها ليس لها بناء أو نمط أو شكل ثابت، كما أنها لا تخضع لأي قواعد في الإعداد أو الترتيب أو البناء. وتتضمن البيانات غير المهيكلة البيانات التي ترد في النصوص وملفات الفيديو، الرسائل الإلكترونية،

العروض التقديمية، التعليقات على صفحات التواصل الاجتماعي، الصور.. الخ. فعلى سبيل المثال أي صفحة ويب يتم إعدادها بلغة HTML تعد مثالاً واضحاً للبيانات غير المهيكلة. وعادةً ما يكون من الصعب تخزين هذه النوعية من البيانات في قاعدة بيانات مهيكلة، إلا إذا تم وضعها ككيانات ثنائية كبرى (Binary Large Objects (BLOBS، وعلى الرغم من أن البيانات غير المهيكلة قد يكون لها في بعض الأحيان شبه هيكل أو بنية كما هو الحال في رسائل البريد الإلكتروني التي يكون لها عنوان مرسل ومستقبل، وموضوع.. إلخ، كما أن صفحات الويب أيضاً تشتمل على مجموعة من الأكواد المحددة مسبقاً، إلا أن المعلومات لا يتم تخزينها سواء في جسم رسالة البريد الإلكتروني أو في متن صفحة الويب بطريقة يمكن من خلالها تصنيف المعلومات بشكل يشبه النماذج الإلكترونية أو قواعد البيانات المهيكلة.

4.1.2 البيانات شبه المهيكلة Simi Structured Dta

تقع تلك النوعية من البيانات في منطقة وسط بين البيانات المهيكلة والبيانات غير المهيكلة. وهي بيانات منتظمة إلى حد ما، من حيث المحتوى، ولكنها غير منتظمة في هيكلها بصورة كاملة وصارمة، كما هو الحال في البيانات المهيكلة. وتشتمل على بيانات غير منتظمة يتم ترتيبها وفقاً لأساليب بناء محددة مسبقاً، ما يساعد على وصفها وفقاً لخصائص محددة تسمح بالبحث فيها باستخدام آليات عامة ولخدمة أغراض عامة.

وعادةً ما يتم تنظيم البيانات شبه المهيكلة في صورة كيانات، بحيث يتم تجميع الكيانات المتشابهة معاً، إلا أنه ليس شرطاً أن تحمل نفس الكيانات محددات متشابهة، كما أنه ليس من الضروري أن يتم ترتيب محددات البيانات في نفس المجموعة أو الحقول.

ومن أبرز أمثلة البيانات شبه المهيكلة السير الذاتية التي لا يوجد لها شكل نمطي أو معياري. فمن الممكن أن يبدأ أحد الأشخاص سيرته الذاتية بعرض الوظائف

السابقة التي شغلها، ثم يعرض الشهادات التي حصل عليه، ثم الأبحاث التي قام بها. ويمكن لشخص آخر أن يبدأ سيرته الذاتية بالشهادات التي حصل عليها، ثم يعرض الوظائف التي شغلها، ثم يعرض المهارات والخبرات، ولا يخصص جزءاً للأبحاث والدراسات، بينما يهتم الأول بوضع جزء خاص للبحوث والدراسات. من هنا يمكن القول إن البيانات شبه المهيكلة عادة ما تضع البيانات في عناصر بيانات دون تحديد صارم لمحتوى وهيكل وترتيب البيانات.

وتعد لغة التكويد الموسعة[1] XML أبرز وسيلة لوضع البيانات شبه المهيكلة في صورة نمطية، حيث إنها معيار واقعي (مصطنع) Defacto يستخدم في وصف الوثائق المتفقة في بعض العناصر وفي شكل البناء، ما يجعل منها نموذجاً دولياً لتبادل البيانات على الويب وبين مؤسسات الأعمال. وتدعم لغة التكويد الموسعة عملية بناء وتطوير الوثائق شبه المهيكلة، والتي تشتمل على كل من بيانات الميتاداتا والنصوص ذات الشكل شبه النمطي.

ويتم تحديد بيانات الميتاداتا باستخدام أكواد لغة التكويد الموسعة. من ثم فإن لغة XML توفر طريقة واضحة وظاهرة لمعالجة البيانات شبه المهيكلة، حيث تعتمد تلك اللغة على محدد نوع الوثائق[2] DTD أو[3] XSD كنماذج لتعريف البيانات شبه المهيلكة وعرضها باستخدام اللغة.

4.1.3 البيانات المهيكلة
Structured Data

البيانات المهيكلة هي عبارة عن بنيات صارمة من حيث الشكل والحجم، ويتم وصف كياناتها بمحددات ثابتة ومحددة، ويتم تنظيمها في صورة تسجيلات

---

(1) XML: eXtensible Mark Up Language

(2) محدد نوع الوثيقة DTD – Document Type Defination

(3) (XSD) XML Schema Definition

يتم تخزينها في جداول بقواعد البيانات. وتتشابه كل التسجيلات التي تتضمنها البيانات المهيكلة في حقول البيانات التي تستخدم في وصفها، ويتم تجميع وتنظيم البيانات في صورة كيانات تساعد على تجميع البيانات المتشابهة في مجموعات باستخدام العلاقات Relations والأقسام Classes. وتحمل الكيانات المتشابهة في نفس المحددات بحيث تتشابه كل الكيانات التي تتضمنها منظومة وصف البيانات Scheme في شكل البيانات، ويكون لها طول محدد مسبقاً وتتبع ترتيباً موحداً. وتعد البيانات المهيكلة من أوائل أنواع البيانات التي تم استخدام الحاسوب في معجالتها.

وقد تم تطوير قواعد البيانات العلائقية لبناء مستودعات بتلك النوعية من البيانات منذ المراحل الأولى لميكنة العمل في المؤسسات. وفي الآونة الأخيرة بدأت أنظمة أكثر تطوراً مثل إدارة علاقات العملاء Customer Relationship management وتخطيط موارد الشركات Enterprise Resource Planing (ERP) ونظم إدارة المحتوى Content management system (CMS) تعتمد على البيانات المهيكلة كنموذج أساسي لمعالجة بياناتها.

وتجدر الإشارة إلى أن عملية تمثيل البيانات في نظام استرجاع المعلومات تتعامل مع ثلاثة أنواع أساسية من البيانات وهي: الميتاداتا بأنواعها المختلفة والنصوص الكاملة، والوسائط المتعددة. وفيما يلي سيتم مناقشة آلية التعامل مع كل نوع من هذه الأنواع والتحديات التي تواجه عملية التمثيل والحلول المتاحة لذلك.

4.2 الميتاداتا Metadata

تم صك مصطلح الميتاداتا لأول مرة في عام 1990 للإشارة إلى عمليات وصف المعلومات الرقمية المتاحة من خلال شبكة الإنترنت، ما أدى إلى ظهور العديد من معايير الميتاداتا التي تم تطبيقها في تمثيل وتنظيم مصادر المعلومات المتشابكة. ثم توسع استخدام المصطلح بصورة كبيرة ليشمل كل ممارسات تمثيل وتنظيم المعلومات، خاصة مع زيادة الاعتماد على شبكة الإنترنت حتى أضحت المنصة

الرئيسة لإنتاج وتمثيل وتنظيم وإتاحة المعلومات الرقمية منذ نهايات القرن العشرين (عبد الهادي، محمد، 2015).

4.2.1 مفهوم الميتاداتا

يمكن تعريف الميتاداتا بأسلوبين مختلفين؛ الأول ضيق في مجاله، حيث يركز على المعلومات الرقمية ويشير إلى وصف مصادر المعلومات الرقمية والمتشابكة باستخدام نموذج معياري مثل معيار (دبلن المحوري Dublin core) والذي تم إعداده خصيصاً لهذا الغرض. والتعريف الآخر أوسع في تغطيته، حيث يشمل كل عمليات تنظيم المعلومات (الفهرسة، التكشيف، التقسيم إلى فئات.. الخ)، والتي يتم إعدادها لأي نوع من أنواع الوثائق سواء بالطرق التقليدية أو غير التقليدية. وفي هذا السياق يمكن النظر إلى بيانات الفهرسة التي يتم إعدادها باستخدام قواعد الفهرسة مثل قواعد الفهرسة الأنجلوأمريكية أو قواعد وصف وإتاحة المصادر أو خطة تصنيف ديوي العشري أو الفهرسة المقروءة آلياً باستخدام شكل الاتصال MARC (Machine Readable catalog) على أنها جميعها نظم ميتاداتا (عبدالهادي & محمد، 2015).

ومن الممكن أن يتم إعداد بيانات الميتاداتا من خلال المؤلف أو منشئ الوثيقة أو أخصائي الميتاداتا أو مدير المستودع أو جهة خارجية تعمل كطرف ثالث Third Party (Dempsey & Heery,1998))، وأحياناً يتم زرع بيانات الميتاداتا في صفحات الويب باستخدام أكواد لغة النصوص الفائقة Hypettext Markup Languege - HTML. ويرى وول (Wool,1998) أنه على الرغم من أن الميتاداتا تتيح نموذجاً فعالاً لوصف وتمثيل المعلومات الرقمية المتاحة في بيئة الإنترنت، إضافة إلى الأنظمة التقليدية مثل التصنيف والفهرسة والتكشيف؛ إلا أنها في الحقيقة امتداد لهذه الأنظمة التقليدية. فكما أشرنا سابقاً إلى أن الطرق التقليدية لا تصلح لتنظيم مصادر المعلومات الرقمية المتاحة على الإنترنت، نظراً للملامح الخاصة التي تتميز بها تلك المصادر والتي سيتم عرضها في الجزء التالي.

4.2.2 ملامح مصادر المعلومات الرقمية المتاحة على الإنترنت

تتميز مصادر المعلومات الرقمية بمجموعة من الملامح الخاصة التي تميزها عن المصادر المطبوعة تشمل (محمد، 2013):

- أنها تتطلب توافر تجهيزات خاصة تشمل المكونات المادية والبرمجيات اللازمة لعرض المحتوى الرقمي.
- أن الشكل Format الذي يتم تسجيل المعلومات الرقمية عليه يتغير بصفة دائمة كنتيجة لسرعة تحديث المكونات المادية والبرمجية، ما يتطلب معه إجراء تهجير للبيانات Data Migration من الشكل القديم إلى الأشكال الحديثة، حيث إنه كثيراً ما يحدث عدم توافق بين الإصدارات المختلفة لنفس البرنامج، وتصبح قضية التوافق أكثر سوءاً عندما يتم تجميع المعلومات الرقمية باستخدام برنامج لتجميع النصوص وآخر للأشكال والجداول وثالث للصور.. إلخ.
- يتم بناء مصادر المعلومات الرقمية باستخدام نمط البناء المعتمد على الهيكل فائق الربط Hyper structure والذي يختلف تماماً عن البناء المسطح Flat Structure للمصادر المطبوعة، ما يجعل من نمو المعلومات وترابطها أمراً من الصعب التحكم فيه. وقد ساعد التقدم الكبير في تطبيقات الإنترنت على تيسير عمليات التواصل والمشاركة بين البشر، لكن ذلك نتج عنه عدم وجود منظومة محكمة لضبط جودة المعلومات والذي ينتج عن الفيضان الهائل من المعلومات المتنوعة من حيث مدى جودتها وإمكانية الاعتماد عليها. لذلك يجب تطبيق طرق متنوعة لتنظيم وتمثيل مصادر المعلومات الرقمية تتوافق مع طبيعة تلك المصادر، حيث إن الأساليب التقليدية وخطط التصنيف وقواعد الفهرسة الأنجلوأمريكية، والفهرسة المقروءة آلياً، تم تطويرها قبل ظهور هذا الكم الهائل من المعلومات الرقمية وتم تصميمها في الأساس لوصف وتمثيل مصادر المعلومات المطبوعة. ومن ثم يمكن القول إن الميتاداتا تم تطويرها لكي تحل مشكلة تمثيل مصادر المعلومات الرقمية التي

يتم تصميمها بالاعتماد على الربط الفائق، ويتم تغيير محتواها بصفة دائمة، إضافة إلى أنها غير متوافقة في جودتها وهائلة في حجمها.

4.2.3 نماذج لمعايير الميتاداتا

على الرغم من أن مصطلح المتياداتا هو مصطلح جديد في مجال تمثيل المعلومات، فقد تم تطوير عدد كبير من معايير الميتاداتا منذ نهاية القرن الماضي وجارٍ تطوير غيرها من المعايير، ويُعد كل من معيار دبلن المحوري وإطار وصف المصادر Resourse Description framework (RDF) أهم النماذج المستخدمة في هذا الإطار (عبدالهادي & محمد، 2015).

وكما أشرنا من قبل، نشأت معايير الميتاداتا أساساً بغرض وصف وتنظيم المعلومات في البيئة الرقمية. ومع الأخذ في الاعتبار طبيعة مصادر المعلومات الرقمية ومصادر الإنترنت. وفي هذا الإطار توجد مجموعة من التساؤلات الأساسية التي تحتاج إلى إجابات واضحة هي كالتالي:

4.2.4 أهمية الميتاداتا في البيئة الرقمية؟

تعتمد عمليات تمثيل المعرفة في البيئة التقليدية لمصادر المعلومات المطبوعة على أعداد تسجيلة ببليوغرافية تشتمل على عناصر الوصف لكل مصدر من مصادر المعلومات، سواء كان بمجموعات مكتبة معينة أو بقاعدة بيانات. ويتم تنظيم تلك التسجيلات كبدائل لمصادر المعلومات تستخدم في عمليات البحث والاسترجاع، إلا أن الممارسة نفسها غير قابلة للتطبيق مع المعلومات الرقمية المتاحة على الإنترنت للأسباب السابق ذكرها، لذلك ظهرت مجموعة من التساؤلات تتعلق بتمثيل وتنظيم مصادر المعلومات الرقمية.

السؤال الأول يتعلق بشكل التمثيل والقواعد التي يتم استخدامها في عمليات التنظيم والوصف، وحيث إن إعداد بديل تقليدي للمصدر الرقمي، كما هو الحال في المصادر المطبوعة لم يعد حلاً مناسباً، فما هو الشكل الملائم لتمثيل المصادر الرقمية والقواعد التي يجب تطبيقها؟

كما ظهر سؤال آخر مرتبط بالمشكلة نفسها، وهو من سيقوم بإنشاء الميتاداتا؟ مع الوضع في الاعتبار حجم المصادر ووحدات المعلومات والكم الهائل المتاح في البيئة الرقمية وخاصة الإنترنت. فالعمر الافتراضي للمصدر الرقمي يعتمد بصورة كبيرة على إتاحه وتوافر التكنولوجيا اللازمة لتشغيله، سواء كانت مكونات مادية أو برمجية والمستخدمة في إنشائه أو إتاحته.

والسؤال الثالث في هذا الإطار مرتبط بالتطور السريع لتكنولوجيا المعلومات الذي يصحبه ضرورة التأكد من أن المصدر الرقمي بمجرد وصفه يمكن الوصول إليه واسترجاعه خلال العمر المتوقع له، ففي بيئة مصادر المعلومات المطبوعة، يظل المحتوى ثابتاً دون تغيير، وأي تغيير يأخذ شكل إصدارة جديدة. أما في البيئة الرقمية فإن المحتوى الخاص بكل وثيقة من الممكن تغييره وبشكل دائم، من ثم لا يمكن التمييز بين الإصدارات المختلفة، بالتالي كيف يمكن التعامل مع الطبيعة الديناميكية لتلك المصادر عند إعداد الميتاداتا الخاصة بها؟

وكما ذكرنا سالفاً، يوجد العديد من معايير الميتاداتا التي تستخدم في تمثيل الكيانات الرقمية في بيئة الإنترنت، وفي الوقت نفسه توجد المعايير التقليدية التي تم استخدامها في تمثيل المعلومات عبر العصور مثل قواعد الفهرسة ونظم التحليل الموضوعي والتصنيف العشري وشكل الاتصال مارك. بالتالي كيف يمكن تضمين معايير الميتاداتا مع غيرها من معايير الوصف سابقة الذكر.

وقد ناقش كل من ديمبسي وهيري (Dempsey & Heery,1998) هذه القضية وأشارا إلى أن مجتمع المعلومات يسعى إلى تحقيق التكامل بين البيئة التقليدية والبيئة الرقمية من خلال ابتكار معايير أكثر شمولاً تستطيع الربط بين المصادر في البيئتين. ولعل إحدى هذه المحاولات هي تجربة شبكة[1] OCLC 's لتطوير نظام ديوي العشري باستخدام أداة مثل Wordsmith والتي تقوم باشتقاق المفاهيم الجديدة والمستجدة والمصطلحات الناشئة من النصوص الرقمية وربطها بخطة تصنيف ديوي العشري (Vizine - Goetz,1997). كما أن

(1) OCLC – Online Computer Library Center

أبرز جهود التطوير في هذا الاتجاه هو تطوير معايير وصف المصادر وإتاحتها Resrource Description and Access التي تسعى إلى وضع آلية وصف جديدة لمصادر المعلومات تراعي متطلبات الوصف في البيئة الرقمية والتقليدية على حد سواء، كما تراعي متطلبات الربط بين مصادر المعلومات بصفة عامة (Wang,2007).

وإضافة إلى كل ما ذكر سابقاً، تبقى قضية التشغيل التبادلي إحدى أهم القضايا التي تحظى بالاهتمام في الوقت الحالي (Rowley & Hartely,(2008. ويشير التشغيل التبادلي إلى قدرة أكثر من نظام؛ لكل منها منصته وواجهة مستفيدين وبنية وهيكل بيانات خاص به، على تبادل ومشاركة البيانات بأقل درجة ممكنة من فقدان المحتوى أو ضعف الأداء الوظيفي (ZHANG, 1998).

وقد ناقش كل من زينج وتشان (Zeng & Chan, 2004) قضية بناء أدوات التشغيل التبادلي بنظم إدارة المعرفة التي عادة ما تستخدم معايير ميتاداتا متنوعة. ومن الواضح أنه ليس من السهل تحقيق التشغيل التبادلي، على الرغم من الجهود الكبيرة التي بذلت في هذا الاتجاه. علاوة على ذلك، فإن كل معيار من معايير الميتاداتا له ملامحه الخاصة وقضاياه المستقلة. فعلى سبيل المثال عند التعامل مع محدد الكيان الرقمي يجب الإجابة على التساؤلات التالية: ما معايير الميتاداتا الذي يجب استخدامها عند تخصيص المحدد؟ هل يجب تحديد أكثر من محدد كيان رقمي لكل شكل جديد أو إصدارة جديدة من نفس العمل؟..إلخ.

وتجد الإشارة إلى أن الأسئلة التي تم طرحها هنا ليست بأي شكل من الأشكال شاملة لكل التحديات التي نواجهها عند التعامل مع قضية تمثيل البيانات الرقمية في بيئة الإنترنت بالاعتماد على معايير الميتاداتا. كما أنه لا توجد خطة للتعامل مع تلك التساؤلات والاهتمامات وتوجد العديد من الممارسات الجديدة في تطبيق وإعداد الميتاداتا للمصادر الرقمية مثل الوصف الانتقائي Selective Description، جداول التحديث والأرشفة المخططة Planned Archiving. ومع ذلك يمكن القول إن الميتاداتا رغم كل ما أثير من تساؤلات حول الممارسات الحالية أو المستقبلية التي يمكن أن تتغير، إلا أنها الطريقة المثلى لتمثيل الكيانات الرقمية، والتي تيسر عملية استرجاعها بكفاءة وفاعلية.

◄ 4.3 النصوص الكاملة
Full Text

يعد تكشيف النصوص الكاملة وإتاحتها للبحث والاسترجاع أحد أهم أهداف نظم تمثيل واسترجاع المعلومات. وقد واجهت عمليات تكشيف النصوص الكاملة صعوبات عدة مع بدايات تطبيق الحاسبات في بناء وتطوير نظم النصوص الكاملة لعل أبرزها: الكُلفة الباهظة لكل من مساحات التخزين ووقت التشغيل اللازمين للتعامل مع الكم الكبير من المعلومات التي يتم تخزينها ومعالجتها. وقد اعتمدت معظم النظم في بداياتها على توظيف بدائل النصوص الكاملة المتمثلة في قواعد البيانات الببليوجرافية والكشافات، بحيث يمكن إتاحة تلك المواد لأغراض البحث والاسترجاع. أما اليوم فقد أصبح من الممكن الاعتماد على جهاز حاسب شخصي في تخزين النصوص الكاملة بسهولة ومعالجتها بسرعة فائقة، لم يَعد ذلك رفاهية في البيئة الرقمية، بل أصبح ضرورة ملحة مع النمو السريع في حجم المعلومات الرقمية التي يتم إنتاجها يومياً، وضرورة إتاحتها للبحث الآني.

◄ 4.3.1 تمثيل معلومات النصوص الكاملة

أدى التطور الملموس في عمليات التخزين الرقمي إلى تحسن كبير في مستويات معالجة النصوص الكاملة (Meadow, et, el,1992). وعلى الرغم من ذلك فإن تمثيل معلومات النصوص الكاملة لتيسير عملية الاسترجاع لا يحتاج إلى «واصف Descriptor لكل كلمة»، ولا تكشيف أو بناء كشاف (Fugmann,1993) سواء كان غير مرئي أو كتمثيل للنص الكامل نفسه. فعملية تمثيل النصوص الكاملة تشبه في خصائصها عملية التكشيف الاشتقاقي من خلال توظيف قوائم الكلمات المستبعدة Stop Lists وجذع الكلمات Stemming وغيرها من التقنيات والآليات المشابهة. وقد وصف لوهان (Luhn, 1960) عملية تمثيل النصوص الكاملة بأنها عملية تكشيف الكلمات المفتاحية وتتم بصورة آلية. وتعتمد معظم نظم الاسترجاع الشهيرة المتاحة على الإنترنت، مثل جوجل وغيره من المحركات، على أسلوب تكشيف الكلمات

المفتاحية، وذلك لتمثيل النصوص الكاملة التي يتم تجميعها في قواعد بياناتهم.

من ثم فإن تمثيل النصوص الكاملة لإتاحتها للبحث والاسترجاع يعد أحد الأساليب الأساسية لتمثيل المعلومات بقواعد بيانات النصوص الكاملة، ولكي تتم تلك العملية لابد أن يتميز محرك البحث بوجود أداة تكشيف للنصوص تستطيع التعرف إلى الكلمات المفتاحية المهمة الواردة بالمادة التي يتم تكشيفها بالاعتماد على خوارزميات معينة وقوائم للكلمات التي يتم استبعادها من عمليات التكشيف.

4.3.2 صعوبات تمثيل النصوص الكاملة

على الرغم من المزايا العديدة التي يمثلها تكشيف النصوص الكاملة من وجهة نظر المستفيد، فإن الناتج النهائي عادة ما يكون معقداً وضخماً، ما يؤدي إلى انخفاض معدلات الاستدعاء، والذي يشير إلى عدد النتائج الصالحة المسترجعة في مقابل عدد النتائج الصالحة في النظام بأكمله. ولعل أبرز مثال على ذلك، حجم النتائج التي يتم استرجاعها من خلال محركات بحث الإنترنت، فعادة ما تسترجع محركات بحث الويب في عملية البحث الواحدة على الأقل عدة آلاف من المواقع يصلح منها عدد محدود جداً للإجابة عن استفسار المستفيد. وقد أشار فوجمان (Fugmann, 1993 P 99) في هذا السياق إلى «أن عمليات تخزين النصوص الكاملة تحتاج إلى مساحات تخزين كبيرة وتستغرق وقتاً طويلاً عند إجراء البحث، ولا يقتصر وقت البحث على استهلاك وقت من جانب الآلات المستخدمة في البحث، ولكن أيضاً يتطلب صبراً من جانب المستفيد لمعالجة الكم الكبير من النتائج المسترجعة».

فاسترجاع النصوص الكاملة، كما سنوضح لاحقاً، هو أحد نماذج تمثيل واسترجاع المعلومات الذي تطور بفضل التطور التكنولوجي الهائل. وعلى الرغم من ذلك فإن جودة عمليات التمثيل والاسترجاع للنصوص الكاملة لاتزال غير مرضية، يظهر ذلك بوضوح في حجم النتائج غير الدقيقة التي يتم استرجاعها من خلال محركات بحث الإنترنت.

وتعتمد الحلول المستقبلية للتغلب على تلك المشكلات على التطور في مجال أبحاث معالجة اللغة الطبيعية، وخاصة التطور في مجالات الويب الدلالي والذكاء الاصطناعي. ومن

المهام الأساسية التي يجب أن تعمل تلك الأبحاث على تحقيقها ما يلي (Knight,1999):

- التطوير في عمليات إعراب الجمل الذي يساعد على تحديد البناء الدلالي للجمل والعبارات.
- اكتشاف وتحديد الكلمات التي تختلف معانيها بحسب ورودها في السياق.

ومن المعروف أن نظم استرجاع المعلومات التي تعمل بصورة آلية لا تتعامل مع المعلومات غير النصية مثل الأشكال والجداول (Fugmann, 1993). لذلك اهتم قطاع من الباحثين والشركات بكيفية معالجة معلومات الوسائط المتعددة مثل الصوت والصور المتحركة المتاحة في صورة رقمية. وسوف يتم استعراض ذلك في الجزء التالي.

4.4 تمثيل معلومات الوسائط المتعددة

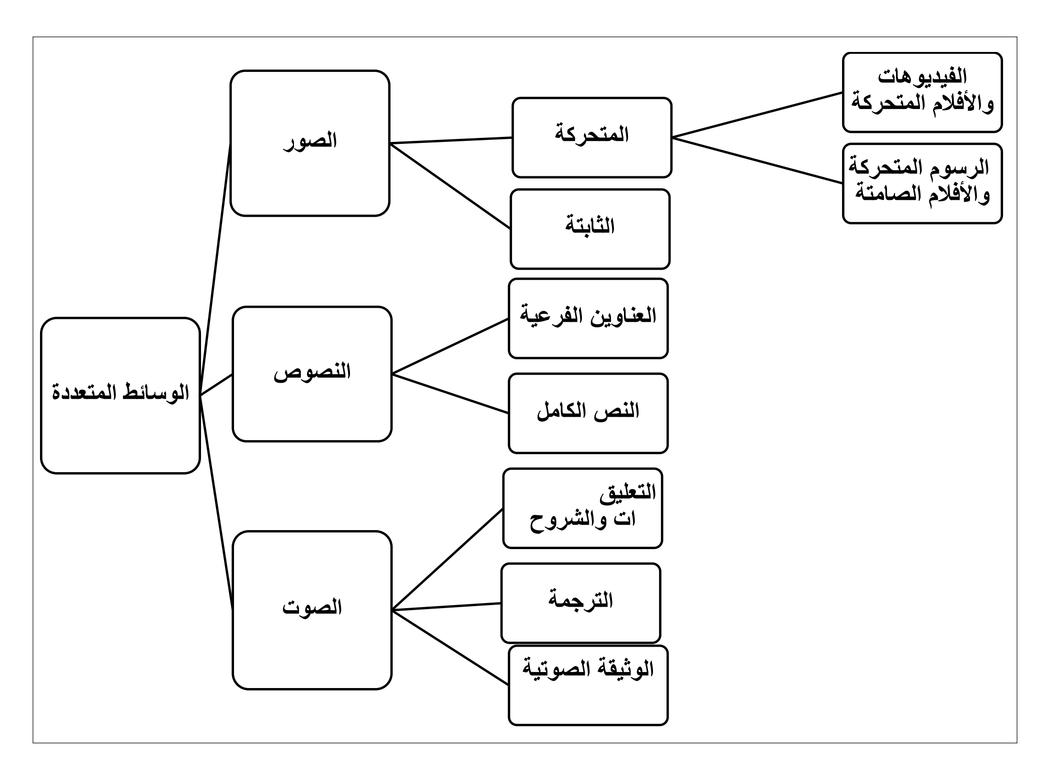
يوجد نمو هائل في حجم معلومات الوسائط المتعددة في البيئة الرقمية، حيث أدى التطور الكبير في آليات إنتاج المعلومات على الشبكة العنكبوتية العالمية إلى تيسير إتاحة تلك النوعية من المعلومات عن ذي قبل. كما أدى ازدهار أساليب إتاحة الوسائط المتعددة على الويب إلى ظهور تحديات كبيرة وجديدة لمجال تمثيل واسترجاع المعلومات.

4.4.1 أنواع معلومات الوسائط المتعددة

الوسائط المتعددة هي أي مزيج من الصوت والصور والمعلومات النصية، سواء كانت الصور ثابتة أو متحركة. وعادة ما يتم استخدام مصطلحي الصوت Sound والمواد المسموعة Audio كمترادفين، وأحياناً ما يستخدم المصطلح وثيقة منطوقة Spoken Document للإشارة إلى المعلومات النصية المسجلة (مثل الخطابات والمحادثات) والتي يطلق عليها الآن المواد المسموعة. وفيما يتعلق بمعلومات

الصـور، فـإن الصـور الثابتـة تشـير إلـى الرسـومات والصـور الفوتوغرافيـة والملصقـات Posters.. . إلـخ، والصـور المتحركـة التي قـد تمتـزج أو لا تمتـزج بالصـوت. ويُشـار إلـى الصـور المتحركـة التي لا تشـتمل على صـوت بالرسـوم المتحركـة Animations أو الأفـلام الصامتـة Silent Movies. ويُطلـق مصطلـح الوسـائط المتعـددة علـى الصـور التـي تمتـزج بالصـوت (الأفـلام أو الفيديوهـات)، ومـن الممكـن أن يظهـر الصـوت ممزوجـاً بالنـص كتعليقـات علـى الصـور Annotation أو ترجمـة، كمـا يظهـر النـص فـي الصـور كشـرح Caption أو عناويـن فرعيـة Subtitles.

ويوضـح الشـكل رقـم (4.1) تشـريحاً تفصيلياً لأنـواع مصـادر المعلومـات المتاحـة فـي صـورة وسـائط متعـددة:



شكل رقم (4.1) تشريح لمصادر معلومات الوسائط المتعددة

4.4.2 أساليب تمثيل الوسائط المتعددة

اعتمد تمثيل الوسائط المتعدده في الماضي، على أساليب الوصف التقليدية التي تستند إلى أساليب الفهرسة الوصفية مثل اسم المنشئ، حجم الصورة، التعليقات والعناوين الفرعية والكلمات المفتاحية..الخ. وقد كان هذا الأسلوبَ الأساسي المستخدم في فهرسة المواد السمعية والبصرية بالمكتبات ومؤسسات المعلومات. وعلى الرغم من أن عملية تمثيل الوسائط المتعددة تعتمد دائماً على التدخل البشري، إلا أنه مازال هناك قصور في جودة المنتج النهائي. ومن بين الأسباب التي تؤدي إلى ذلك أنه مازال من الصعب وصف الوسائط المتعددة بصورة صريحة وموضوعية. فعلى سبيل المثال كيف يمكن وصف صورة شروق الشمس أو غروبها، أو قطعة موسيقية هادئة أو حتى صاخبة باستخدام مصطلحات تعبر عن محتواها بشكل صريح، إضافة إلى ذلك كيف يمكن تحقيق الاطراد والدقة في عملية التمثيل لمعلومات الوسائط المتعددة بالاعتماد على التوجه غير الموضوعي الذي يتضمن قدراً كبيراً من الذاتية والآراء الشخصية.

لقد تم تطوير أسلوب التمثيل المستند إلى المحتوى Approach Content Based لتمثيل الوسائط المتعددة من خلال خصائصها مثل لون الصورة، النغمات الصوتية، وذلك للتغلب على القصور والقيود التي يفرضها الأسلوب المستند إلى الوصف Description Based Approach السابق عرضه. واعتمد تطوير آليات التمثيل المستند إلى المحتوى من خلال تطوير تقنيات تستطيع وصف المحتوى مثل التعرف الصوتي Speech Recognition والتعرف النمطي Pattren Recognition وفهم الصور Image Understanding والتي تستخدم في وصف وتحليل الوسائط المتعددة لأغراض التمثيل.

ويعد هذا التوجه رمزاً لتغيير نماذج تمثيل الوسائط المتعددة، فإذا كان نموذج التمثيل المستند إلى وصف الوسائط المتعددة يتم إنجازه من خلال المعلومات الوصفية وبطريقة يدوية مثل المنشئ، وسنة الإنتاج والحجم..، ومعلومات المحتوى من خلال (الكلمات المفتاحية ورؤوس الموضوعات)، فإن التمثيل المستند إلى المحتوى يعمتد على تحليل خصائص ومحددات الوسائط المتعدده مثل ألوان الصور، النغمات الصوتية.. إلخ.

وتشتمل خصائص الوسائط المتعددة على أوجه متنوعة، لعل أبرزها الخصائص المشتركة للصور الثابتة مثلاً؛ اللون، الشكل، النصوص، والتي يمكن تفصيلها وتحليلها أكثر من خلال خصائص مثل الاتجاهية Directionality العشوائية Randomness، التماسك Robustness التضاد Contrast وغيرها.

أما المعلومات الصوتية فيمكن تحليل خصائصها إلى مجموعة من المعاملات تشمل السرعة والنغمات والنترات، بحيث يمكن استخدامها في عمليات التمثيل. وتمثل هذه الملامح الأساسية عن الوسائط المتعدده الحد الأدنى من المعلومات التي يمكن اشتقاقها آلياً أو بطريقة شبه آلية، والتي تحد أو تقلل بقدر كبير من الحاجة إلى التدخل البشري الذي مازال مكلفاً وغير مرغوب في عملية تمثيل الوسائط المتعددة بدرجة كبيرة.

وتعتمد آليات تمثيل الفيديو والصور والرسوم المتحركة على مجموعة من الخصائص تشبه تمثيل الصور الثابتة والأصوات، إلى جانب اتخاذ إجراءات التقطيع أو التجزئة للملف Segmentation . وقد قامت شو (Chu 2001) بالمقارنة بين هذين الأسلوبين لتحديد أيهما أكثر استخداماً في البحوث والتطبيقات، حيث قامت بتحليل الاستشهادات المرجعية للإنتاج الفكري المنشور في مجال تكشيف واسترجاع الصور، وتوصلت إلى أن التمثيل المستند إلى المحتوى قد سيطر على الدراسات والتطبيقات في هذا المجال في السنوات الأخيرة. وقد أشارت شو إلى أن السبب الرئيس وراء ذلك هو التعقيد الذي يتضمنه تطبيق أسلوب التمثيل المستند إلى الوصف في مقابل التطور التكنولوجي الهائل في آليات دعم التمثيل المستند إلى المحتوى، الذي أدى بدوره إلى تيسير عمليات التحليل واستخلاص المعلومات الدالة على المحتوى. ومع ذلك فإن نتائج شو تشير إلى أن تمثيل الوسائط المتعددة المستند إلى المحتوى لا يمكن أن يمثل الأسلوب الوحيد في المستقبل، على العكس من ذلك فإن أسلوب التمثيل المستند إلى الوصف إذا تم تطبيقه بطريقة فعالة (أقل كُلفة) وأكثر اطراداً وموضوعية فإنه قد يساعد بصورة كبيرة على تحقيق الجودة في تمثيل معلومات الوسائط المتعددة لذلك فإن النموذج الأمثل هو المزج والتكامل بين الأسلوبين في تمثيل الوسائط المتعددة.

4.4.3 تحديات تمثيل الوسائط المتعددة

إلى جانب ما سبق ذكره من مشكلات مرتبطة بتمثيل الوسائط المتعددة، فإن مشكلات تجزئة الصور المتحركة Moving Image Segmentation وتحليل الخطب والمحادثات Speech Parsing أو الصوت مازالت تمثل تحديات أساسية في مجال تمثيل الوسائط المتعددة. فعملية تجزئة الصور المتحركة تعد خطوة أساسية نحو فك الصور المتحركة إلى وحدات (مثل تشغيل الكاميرا، لحظات الصمت) بمعنى الفواصل بين عناصر العمل.

كما يتم تحليل الكادرات الأساسية Key Frames التي تشمل الكادرات التي تتضمنها كل لقطة Shot والتي يتم استخدامها كأساس لتحليل المحتوى وتمثيله (Zhang, et. al., 1995). وتوجد أساليب متنوعة تشمل تقنيات وخوارزميات لتجزئة الصور المتحركة، ويظل جوهر تلك العملية واحداً في كل تلك الأساليب، حيث يعتمد على تقسيم الصور المتحركة إلى كيانات صغيرة تحمل دلالات من ثم يمكن تحليلها وتمثيلها بتساوٍ وتوازن ودقة.

وتعتمد عملية تجزئة الخطاب Speech Segmentation على تقطيع الخطاب الكامل إلى فقرات وجمل وعبارات وكلمات، بحيث يمكن تحديد محتواه الموضوعي وتمثيله. ومن الصعب تحديد معايير خاصة بطريقة بناء الصور المتحركة أو المعلومات الصوتية، نظراً للطبيعة الخاصة والمعقدة المرتبطة بهما. ذلك أن الصور المتحركة مستمرة في الزمن والمساحة، ولا يمكن دائماً الاعتماد على الفواصل بين كادرات الكاميرا المتصلة Consecutive Camera Shots وتجزئتها، لأنه أمر صعب، كما أن الخطابات الصوتية لا تتضمن أي علامات ترقيم أو فواصل بين الجمل والكلمات أو غيرها من العلامات التي تساعد على عملية التجزئة، كما هو الحال في المواد النصية المكتوبة، ما يجعل عملية تجزئة المواد المسموعة مهمة صعبة.

ومن أبرز أمثلة الصعوبات التي تواجه عملية التمثيل المواقف والإشارات التي تتضمنها المواد الصوتية مثل لحظات الصمت (أصوات التنفس، تلعثم اللسان،

الهمهمات..إلخ) عدم الطلاقة في الكلام (مثل الكلمات المنفصلة عن أي سياق، التوقفات Pauses، التردد وتغيير الكلمات أو العبارات)، وكلمات إضافة أحداث وأمثلة مثل (وإضافة إلى وعلى سبيل المثال..إلخ)، وبسبب كل هذه التحديات والصعوبات فإن عملية التدخل البشري في الوصف الدقيق للوسائط المتعددة مازالت ضرورة ملحة حتى مع النظم التي تعتمد على التمثيل المستند إلى المحتوى، وما زالت هناك حاجة إلى مزيد من الدراسات والبحوث في هذا الاتجاه بغرض تحقيق الدقة والشمول والجودة في المعالجة والتمثيل.

وتجدر الإشارة بصفة عامة إلى أنه يوجد عدد محدود من الدراسات والبحوث التي تمت على عمليات تمثيل واسترجاع المعلومات غير النصية. وتعد المواد الصوتية أقل أنواع المواد التي حظيت بعناية من بين الأنواع المتعددة للمواد التي تتضمن معلومات وسائط متعددة، في نفس الوقت الذي تشهد هذه النوعية من المصادر نمواً مطرداً في حجم المعلومات وفي عدد الوسائط المتعددة في البيئة الرقمية، والتي أصبحت تمثل نسبة كبيرة من حجم الويب الفعلي، حيث ترى بعض التقديرات أنها تجاوزت نسبة 30 ٪ من حجم الويب (Jones et,el., 1996, and Djeraba,(2002. لذلك فإن التغلب على تحديات معالجة وتمثيل الوسائط المتعددة يمثل ضرورة كبرى لتيسير عمليات تمثيل واسترجاع وإتاحة تلك المعلومات. وتوجد حاجة ماسة إلى إجراء العديد من البحوث والدراسات في هذا المجال للتغلب على التحديات التي تواجه الطرق الآلية لتجزئة المواد الصوتية والصور المتحركة. فمازال التدخل البشري عنصراً مهماً وأساسياً في تمثيل الوسائط المتعددة حتى باستخدام أسلوب التمثيل المستند إلى المحتوى Approach Content Base.

ويمكن القول بصفة عامة إن عدد الدراسات والبحوث التي اهتمت بتمثيل الوسائط المتعددة مقارنة بالمواد النصية مازال محدوداً جداً. وتعد المواد المسموعة أقل المواد التي حظيت بعناية واهتمام الباحثين من بين مواد الوسائط المتعددة. ومع النمو المطرد في عدد الوسائط المتعددة في البيئة الرقمية، فإن تمثيل الوسائط المتعددة يمثل تحدياً حقيقياً لإتاحة المعلومات التي تتضمنها تلك الوسائط.

◄ 4.5 إطار ملخص لتمثيل المعلومات

يُعد تمثيل الوحدات المعرفية عملية أساسية عند استرجاع المعلومات لسببين أساسين هما:

- الأول أن التمثيل يوفر بدائل أكثر فعالية في البحث والاسترجاع لذلك فإن المعلومات لابد أن تكون ممثلة قبل أن يتم استرجاعها.
- جودة التمثيل تؤثر بصورة مباشرة في كفاءة الأداء في عملية الاسترجاع.

تعتمد عملية تمثيل المعلومات لأغراض الاسترجاع على معلومات وصفية مظهرية Offness ومعلومات عن المضمون aboutness باستخدام النموذج المستند إلى المحتوى في تمثيل الوسائط المتعددة. وتجدر الإشارة إلى أن المعلومات المظهرية تشمل خصائص وصفية للمادة التي يتم تمثيلها مثل المؤلف أو المنشأ، اللغة، سنة النشر..إلخ، أما معلومات المضمون فتتعامل مع المحتوى الموضوعي للوثائق والمعلومات. ويعد نموذج المعلومات الوصفية المظهرية في التمثيل أكثر وضوحاً وسهولة مقارنة بنموذج معلومات المضمون الذي يعد أكثر صعوبة وتعقيداً، حيث يعاني من مشكلات معالجة اللغة، والتي سبق عرضها، وخاصة التعامل مع المترادفات والمشترك اللفظي..الخ.

ويتأثر أداء نظام الاسترجاع بكفاءة نظام التمثيل، لذلك لابد من الاهتمام بتحقيق أعلى مستويات الكفاءة والدقة والاطراد في تمثيل المعلومات في البيئة الرقمية، ما سيكون له بالطبع تأثير الكبير في سرعة وسهولة الوصول إلى المعلومات في تلك البيئة المعقدة والمتشابكة.

المصادر

- عبدالهادي، محمد فتحي؛ موسى، خالد عبدالفتاح (2015). الميتاداتا: أسسها النظرية وتطبيقاتها العملية. الدار المصرية اللبنانية، القاهرة، 262 ص.
- محمد، خالد عبدالفتاح (2018). الويب الدلالي وتطبيقاته في المكتبات ومؤسسات المعلومات. الرياض، دار المتنبي، 317 ص.
- محمد، خالد عبدالفتاح (2013). تحليل وفرز النتائج في محركات بحث الشبكة العنكبوتية، دراسات عربية في المكتبات والمعلومات، دار غريب للطباعة والنشر والتوزيع 11 ص ص 8ـ23.

- Aitchison, J., Bawden, D., & Gilchrist, A. (2003). Thesaurus construction and use: a practical manual. Routledge.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 34-43.
- Chen, H., Ng, T. D., Martinez, J., & Schatz, B. R. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: an esxperiment on the worm community system. Journal of the American Society for Information Science, 48(1), 17-31.
- Chu, H. (2001). Research in image indexing and retrieval as reflected in the literature. Journal of the American Society for Information Science and Technology, 52(12), 1011-1018.
- Dempsey, L., & Heery, R. (1998). Metadata: a current view of practice and issues. Journal of documentation, 54(2), 145-172.
- Djeraba, C. (2002). Content-based multimedia indexing and retrieval. IEEE multimedia, 9(2), 18-22.
- Fugmann, Robert. Subject analysis and indexing: theoretical foundation and practical advice. Vol. 1. Indeks Verlag Dr. Ingetraut Dahlberg, 1993.
- Gilchrist, A. (2003). Thesauri, taxonomies and ontologies–an etymological note. Journal of documentation, 59(1), 7-18.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.
- Jones, K. S., Jones, G. J. F., Foote, J. T., & Young, S. J. (1996). Experiments in spoken document retrieval.Information Processing & Management,32(4), 399-417

- Knight, K. (1999). Mining online text. Communications of the ACM, 42(11), 58-61.
- Lancaster, F. W. Vocabulary Control for Information Retrieval, 2d ed.(Arlington, Va.: Information Resources Pr., 1986); Thomas Mann. Library Research Models: A Guide to Classification, Cataloging and Computers, 486-91.
- Lancaster, F. W., & Warner, A. J. (1993). Information Retrieval Today. Revised, Retitled. Information Resources Press, 1110 North Glebe Rd., Suite 550, Arlington, VA 22201..
- Lei Zeng, M., & Mai Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. Journal of the American Society for information science and technology, 55(5), 377-395.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.
- Meadow, C. T., Boyce, B. R., & Kraft, D. H. (1992). Text information retrieval systems (Vol. 20). San Diego, CA: Academic Press.
- Milstead, J. L. (1995). Invisible thesauri: the year 2000. Online and CD-Rom Review, 19(2), 93-94.
- Munk, T. B., & Mork, K. (2007). Folksonomy, the power law & the significance of the least effort. Knowledge organization, 34(1), 16-33.
- National information standards organization. (1993). Guidelines for the construction, format, and management of monolingual thesauri(ANSI/NISO Z39.19-1993). Bethesda, MD: NISO press.
- Noruzi, A. (2006). Folksonomies:(un) controlled vocabulary?. Knowledge organization, 33(4), 199-203.
- Rowley, J. (1992). Organizing knowledge: an introduction to information retrieval. Gower.
- Rowley, J. (1994). The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. Journal of information science, 20(2), 108-118.
- Rowley, J., & Hartley, R. (2017). Organizing knowledge: an introduction to managing access to information. Routledge.
- Schatz, B. R. (1997). Information retrieval in digital libraries: Bringing search to the net. Science, 275(5298), 327-334.

- Speller, E. (2007). Collaborative tagging, folksonomies, distributed classification or ethnoclassification: a literature review. Library Student Journal, 2.
- Spiteri, L. F. (2007). The structure and form of folksonomy tags: The road to the public library catalog. Information technology and libraries, 26(3), 13-25.
- Trant, J., & with the participants in the steve. museum project. (2006). Exploring the potential for social tagging and folksonomy in art museums: Proof of concept. New Review of Hypermedia and Multimedia, 12(1), 83-105.
- Uschold, M. (1996, September). Building ontologies: Towards a uni ed methodology. In Proceedings of 16th Annual Conference of the British Computer Society Specialists Group on Expert Systems.
- Vander Wal, T. (2007). Folksonomy. http://www.vanderwal.net/essays/051130/folksonomy.pdf
- Vickery, B. C. (1997). Ontologies. Journal of information science, 23(4), 277-286.
- Vizine Goetz, D. (1997). From book classification to knowledge organization: improving Internet resource description and discovery. Bulletin of the American Society for Information Science and Technology, 24(1), 24-27.
- Wang, J. (2007). Digital object identifiers and their use in libraries. Serials review, 33(3), 161-164.
- Wellisch, H. H. (1995). Indexing from A to Z, 2nd. New York: HW Wilson.
- Wool, G. (1998). A meditation on metadata. The Serials Librarian, 33(1-2), 167-178.
- Zeng, lei; Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. Journal of the American Society for information science and technology, 55(5), 377-395Zhang, H., Low, C. Y., Smoliar, S. W., & Wu, J. (1995, January). Video parsing, retrieval and browsing: an integrated and content-based solution. In Proceedings of the third ACM international conference on Multimedia (pp. 15-24). ACM
- Zhonghong, W., Chaudhry, A. S., & Khoo, C. (2006). Potential and prospects of taxonomies for content organization. Knowledge organization, 33(3), 160-169.

الفصل الخامس

اللغة في تمثيل
واسترجاع المعلومات

5 مقدمة

تعد اللغة أحد المكونات الرئيسة لأي نظام من نظم المعلومات عامة، وفي نظم تمثيل واسترجاع المعلومات خاصة. ويوجد نوعان أساسيان من اللغات في تمثيل واسترجاع المعلومات هما اللغة الطبيعية واللغة المضبوطة. وتستخدم اللغتان في ترجمة المفاهيم التي تتضمنها الوثائق التي يتم تمثيلها إلى مصطلحات تستخدم في وصف المفاهيم والمحتوى الموضوعي للوثائق. وعلى الرغم من إمكانية الاختيار بينهما، إلا أن السؤال الخاص بأيهما أفضل، مازال محل جدل دائم بين المتخصصين. وقد نتج عن استخدام لغتين للتعبير عن المصطلحات نظامان للتكشيف: هما نظم تكشيف اللغة المقيدة أو المضبوطة ونظم تكشيف اللغة الطبيعية. وتستخدم اللغة في التعبير عن المحتوى الموضوعي للوثائق باستخدام مصطلحات يتم اشتقاقها من أدوات (نظم اللغة المضبوطة) أو من النصوص مباشرة (نظم اللغة الطبيعية) للتعبير عن المفاهيم التي تتناولها تلك الوثائق. وسيتم فيما يلي التعرف إلى طريقة تطبيق كل نوع من هذين النوعين في نظم استرجاع المعلومات.

5.1 نظم تكشيف اللغات المقيدة أو المضبوطة

هي النظم المبنية على الاختيار والصياغة والربط بين المصطلحات التي تعبر عن المحتوى الموضوعي لأوعية المعلومات من خلال الاعتماد على لغات تكشيف معيارية. ويطلق عليها نظم مضبوطة أو مقننة، نظراً لأن التحكم في المصطلحات وطريقة الربط بينها يتم وفقاً لمعايير معينة تحددها لغة التكشيف التي يعتمد عليها النظام. وتنبع الحاجة إلى استخدام لغات مضبوطة في التعبير عن المحتوى الموضوعي للوثائق من طبيعة اللغة بصفة عامة.

وتعد قوائم اللغات المضبوطة نموذجاً بارزاً للغات الاصطناعية، حيث إن مصطلحاتها وبنيتها ودلالاتها محددة ومقيدة في استخدامها (Wellisch, 199). ومن المعرف أن المصطلح الواحد من الممكن أن تكون له معان مختلفة في أكثر من قائمة مصطلحات مضبوطة، حيث إنه عادة ما يكون لكل قائمة توجهها الخاص. لذلك فإن عمليات تجهيز المصطلحات المضبوطة عادة ما تعتمد في بنيتها على التوجه العام للمؤسسة التي تخدمها. ويعتمد اختيار المصطلحات التي يتم تضمنيها اللغات المضبوطة على مبدأين أساسيين هما:

- السند الأدبي Literary Warranty
- سند المستفيد User Warrant

السند الأدبي يشير إلى أن المصطلح الذي يتم اختياره بالقائمة لابد أن يكون له نظير بالإنتاج الفكري المتخصص في المجال، ما يعنى أنه ظهر بأحد مصادر المعلومات الحديثة وبناء عليه يتم إضافته إلى القائمة، بمعنى أن عملية اختيار المصطلحات وإضافتها إلى قوائم رؤوس الموضوعات تستند في الأساس إلى المصطلحات الواردة باللغة الطبيعية في الإنتاج الفكري. من ثم فإن اللغة الطبيعية عادة ما تكون أكثر ثراءً وتنوعاً من اللغة المضبوطة.

وبالمثل، فإن سند المستفيد يشير إلى أن المصطلح الذي يتم اختياره بالقائمة لابد أن يكون تم استخدامه في استفسارات المستفيدين كمصطلح بحثي في الماضي، أو من المتوقع استخدامه في المستقبل في البحث عن الإنتاج الفكري الذي ظهر به المصطلح في مرحلة السند الأدبي.

من ثم فإن بناء قوائم المصطلحات المضبوطة من الممكن أن يعتمد على تحليل محتوى النصوص لاشتقاق الكلمات ثم يتم ضبطها أو تحليل ملفات لوج استفسارات المستفيدين Users Queries Log. ويوجد ثلاث نماذج للغات المضبوطة هي المكانز وقوائم رؤوس الموضوعات وخطط التصنيف، وسيتم فيما يلي عرض كل نموذج من هذه النماذج بشيء من التفصيل.

5.1.1 وظائف اللغة المقيدة

اللغات المقيدة أو المضبوطة تحقق العديد من الوظائف عند استخدامها كأساس لعملية التكشيف منها ما يلي:

1. الاطراد في التكشيف، بمعنى الثبات على مصطلح واحد محدد للدلالة على المفهوم المكشف. من ثم تساعد على تجنب التشتت الموضوعي في مرحلتي التكشيف والبحث.
2. تيسير إجراء عمليات البحث العريضة والشاملة التي تساعد على تجميع المصطلحات المتصلة ببعضها بعضاً دلالياً، وذلك من خلال الاستفادة من إمكانيات البحث الشامل.
3. ضمان التعبير عن جميع المفاهيم المشتركة لفظياً في الهجاء والمختلفة في الدلالة بمصطلحات مختلفة من خلال التبصرات التي توضح مجال المصطلح.
4. اللغات المضبوطة تتمتع بالقدرة على تحقيق مستويات دقة عالية High Precision Rate في مرحلة البحث.

5.1.2 عيوب نظم اللغة المقيدة

ومن أهم عيوب نظم التكشيف التي تعتمد على اللغات المقيدة ما يلي:

1. الكُلفة الباهظة؛ حيث تحتاج هذه النظم إلى خبراء متخصصين في المجالات الموضوعية وعلى دراية دقيقة ببنية لغات التكشيف ومتطلبات عملية التكشيف.
2. تقادم مصطلحات اللغة وعدم قدرتها على متابعة التطورات التي تحدث في الإنتاج الفكري. وتبرز هذه المشكلة بشكل أكثر وضوحاً عند ظهور مصطلح جديد في الإنتاج الفكري، حيث تشير الدراسات إلى أن أي مصطلح قد يستغرق ما بين عامين إلى ثلاثة أعوام حتى يظهر في لغات التكشيف المضبوطة.

5.1.3 أنواع نظم التكشيف المقيدة

تنقسم نظم التكشيف المضبوطة أو المقيدة إلى فئتين أساسيتين هما:

5.1.3.1 نظم تكشيف الربط المسبق
Pre-coordinate Indexing Systems

وهي النظم التي تربط بين المصطلحات في مرحلة التكشيف، بحيث يتم إعداد تراكيب مصطلحات أو رموز تعبر عن المحتوى الموضوعي للوثيقة أو وعاء المعلومات بكافة جوانبه. وتعتمد هذه الطريقة على استخدام أدوات الربط المسبق مثل قوائم رؤوس الموضوعات وخطط التصنيف لكي تظهر في شكل رؤوس مركبة تضم أو تجمع معاً المصطلحات التي تمثل موضوع الوحدة المكشفة. وتعتمد نظم تكشيف الربط المسبق على أداتين أساسيتين هما قوائم رؤوس الموضوعات وخطط التصنيف.

- **قوائم رؤوس الموضوعات**

تعد قوائم رؤوس الموضوعات من أقدم نماذج قوائم المصطلحات المضبوطة التي تم تصميمها لأغراض الربط المسبق واللاحق معاً. وقد كان الربط المسبق النموذج السائد في البناء حتى الأربعينات من القرن الماضي. ويقصد بالربط المسبق دمج المصطلحات من خلال أنظمة التفريع والتركيب قبل عملية التمثيل والاسترجاع.

قوائم رؤوس الموضوعات هي عبارة عن قوائم منهجية بموضوعات المعرفة البشرية مرتبة ترتيباً هجائياً مع بيان العلاقات بين هذه الموضوعات. وتشتمل قوائم رؤوس الموضوعات على ثلاثة أنواع رئيسة من رؤوس الموضوعات هي:

- **رأس الموضوع المفرد**: ويأخذ هذا الرأس شكل كلمة واحدة مثل الإعلام، المكتبات، الحاسوب.. إلخ.

- **رأس الموضوع المركب**: وهو عبارة عن رأس مكون من كلمتين مركبتين مثل استرجاع المعلومات، الحاسب الآلي، التطوير الذاتي، إدارة الإعمال.. الخ.

- **رأس الموضوع المعقد:** وهو عبارة عن رؤوس الموضوعات التي تتضمن أكثر من كلمتين مثل نظم استرجاع المعلومات، النظم الآلية المتكاملة.

فإذا كانت المادة المكشفة تتناول موضوع Internet Retrieval System فإن نظام الربط المسبق يربط بين تلك المصطلحات في قائمة رؤوس الموضوعات من البداية، من ثم يتم استخدام المصطلح بصورته المعقدة في عملية التمثيل، وكذلك في عملية الاسترجاع. لذلك فإن عملية الربط تتم عند بناء المصطلح لأغراض التمثيل، كما تتم بنفس الطريقة في مرحلة الاسترجاع دون تدخل من المكشف أو الباحث، حيث يجب على كل منهما التزام التتابع الخطي المستخدم في عملية بناء المصطلحات عند التمثيل والاسترجاع. ونظراً لأن قوائم رؤوس الموضوعات تتيح إمكانيات الربط المسبق واللاحق (بدرجة أقل)؛ فإنها تتميز بمرونة أكبر من خطط التصنيف؛ ولكنها أقل تحديداً ومرونة من المكانز.

ومن أهم السمات التي تميز قوائم رؤوس الموضوعات وتجعلها أداة من أهم أدوات الربط المسبق، استخدامها لمبدأ التفريعات، حيث إن رؤوس الموضوعات سواء كانت بسيطة أو مركبة أو معقدة تطبق تراكيب عدة سواء كانت وجهية أو جغرافية أو زمنية أو شكلية. بالتالي فإن رأس الموضوع يرد في القائمة، إما مركباً مع كافة الأوجه الممكنة أو توفر القائمة إمكانية تركيبه من الأوجه المختلفة. ومن أمثلة قوائم رؤوس الموضوعات الشهيرة قائمة رؤوس موضوعات الكونغرس، قائمة رؤوس موضوعات سيرز، قائمة رؤوس الموضوعات الطبية، قائمة رؤوس الموضوعات العربية الكبرى.

عادة ما يستخدم مصطلح رؤوس الموضوعات للدلالة على المصطلحات التي تتضمنها قوائم رؤوس الموضوعات، ويتم ترتيب تلك الرؤوس ترتيباً هجائياً. وتعتمد تلك القوائم على شبكة الإحالات في عمليات الإشارة والتحويل. وأهم أنواع تلك الإحالات إحالة انظر See والتي تستخدم للإحالة من المصطلح غير المستخدم إلى المصطلح المستخدم. بينما تستخدم علامة X والتي تعني انظر من See From والتي تحيل المستخدم إلى التعبير المفضل للمصطلح باستخدام الإحالة انظر See.

- **نماذج للإحالات بقوائم رؤوس الموضوعات**

Handicapped المعوق

See انظر

Physically Challenged متحدي الإعاقة

ومن المصطلح متحدي الإعاقة تستخدم إحالة انظر من (X)

Physically Challenged متحدي الإعاقة

X X

Handicapped المعوق

فعلامة X هنا تشير إلى أن مصطلح متحدي الإعاقة هو المصطلح المفضل لهذا المفهوم.

وتستخدم إحالة انظر أيضاً See For وإحالة XX التي تستخدم للدلالة على انظر أيضاً من See Also From وتستخدم إحالة انظر أيضاً للدلالة على العلاقات الشجرية والبينية (المرتبطة) بين رؤوس الموضوعات. وكما هو الحال في إحالة X فإن إحالة XX تحيل المستخدم إلى المصطلح المفضل See Also.

من ثم يمكن القول بصفة عامة إن قوائم رؤوس الموضوعات تستخدم لأغراض التمثيل الاصطلاحي والمفاهيمي في صورة مقيدة بنظم الربط المسبق واللاحق معاً، إلا أنها أقل استخداماً وشيوعاً من المكانز في نظم التمثيل والاسترجاع بنظم المصطلح غير الواحد.

وتُعد قائمة رؤوس موضوعات مكتبة الكونجرس وقائمة رؤوس موضوعات سيرز Sears أبرز نماذج قوائم رؤوس الموضوعات على المستوى العالمي، مع العلم أن قائمة رؤوس موضوعات مكتبة الكونجرس تحولت منذ الطبعة الحادية عشرة إلى النموذج المكنزي في البناء الهرمي للمصطلحات وشبكة الإحالات. وتعتمد في نسختها المتاحة على الويب على نموذج العرض المرئي للبنية الهرمية

للمصطلحات[1]. وقد بدأت قائمة رؤوس موضوعات مكتبة الكونجرس منذ بداية الألفية الجديدة تطبيق معايير ربط البيانات Linked Data من خلال ربط المصطلحات بتطبيقات إطار وصف المصادر Reasourse Description Framwork – RDF ومعايير الميتاداتا المطبقة بالمكتبة.

وعلى المستوى العربي تعد قائمة رؤوس الموضوعات العربية الكبرى لشعبان عبد العزيز خليفة وقائمته للمكتبات المدرسية والعامة والمعروفة بقائمة رؤوس الموضوعات القياسية من أبرز النماذج العربية وأكثرها انتشاراً واستخداماً. وتجدر الإشارة هنا إلى أن قوائم رؤوس الموضوعات العربية مازالت تعتمد على الأساليب التقليدية في بناء المصطلحات والربط بينها والتعبير عن شبكة العلاقات والمصطلحات. وتوجد حاجة ماسة إلى تطوير أدوات جديدة في البيئة العربية تتوافق مع التطورات التي تسير في هذا المجال واحتياجات تمثيل استرجاع المعلومات في البيئة الرقمية.

- **خطط التصنيف**

هي عبارة عن قوائم منهجية بموضوعات المعرفة البشرية مرتبة وفقاً لخطة تصنيف تربط وتجمع الموضوعات وفقاً لعلاقاتها ببعضها بعضاً. وعادة ما تتدرج خطط التصنيف من الموضوعات العامة إلى الموضوعات الأكثر تخصصاً. وتسمح ببناء تراكيب للموضوعات التي تشتمل على أكثر من جانب موضوعي. وتنقسم خطط التصنيف التي يمكن استخدامها في نظم تكشيف الربط المسبق إلى نظم تصنيف حصرية مثل خطة تصنيف مكتبة الكونجرس، نظم تصنيف شبه حصرية مثل خطة تصنيف ديوي العشري، والنظام العشري العالمي، نظم تصنيف وجهية مثل خطة تصنيف رنجاناثان ونظام تصنيف تشارلز كتر.

وتعد خطط التصنيف أقدم نماذج التكويد باستخدام آليات مضبوطة مسبقاً، أي تستخدم نموذج الربط المسبق في تمثيل المفاهيم والموضوعات. ويطلق

---

http://id.loc.gov/authorities/subjects.html (1)

على الوحدات الأساسية لخطة التصنيف الفئات Classes والتي يتم تمثيلها بصورة رقمية أو هجائية أو مزيج منهما معاً. بمعنى أن خطط التصنيف تستخدم الرموز (الرقمية، الهجائية أو مزيجاً منهما مع علامات خاصة) للدلالة على المفاهيم والموضوعات.

ونظراً لأنها أقدم نماذج نظم التمثيل بآليات التكويد المضبوطة، فإن خطط التصنيف شهدت العديد من التطورات المتلاحقة والمراجعة والتحديث خلال الفترة من نهاية القرن التاسع عشر حتى بدايات القرن الواحد والعشرين. وعلى عكس كل من المكانز وقوائم رؤوس الموضوعات اللذين يستخدمان الإطار الطبيعي في التعبير عن المعرفة من خلال آليات التعبير الاصطلاحي أي باستخدام المصطلحات والكلمات، تعتمد خطط التصنيف على إطار اصطناعي للمعرفة يتمثل في تكويد الموضوعات برموز للدلالة عليها. فعلى سبيل المثال تستخدم خطة تصنيف ديوي العشري نموذجاً اصطناعياً للتمثيل الاصطلاحي للمعرفة مكون من 10 فئات أساسية، ثم يتم تقسيم الفئات الأساسية إلى 10 شعب لكل فئة وهكذا في تدرج منطقي هرمي لتمثيل المعرفة في مقابل التدرج الشجري أو العلائقي المستخدم في المكانز والتدرج الهجائي المستخدم في قوائم رؤوس الموضوعات.

وبالنظر إلى التدرج المنطقي للفئات والشعب نلاحظ أنه تدرج هرمي للعلاقات الاصطناعية بين الموضوعات. من ثم نجد أن بعض الموضوعات يمكن عرضها في إطار أكثر عمقاً في البناء الهرمي من موضوعات أخرى. ويتم التعبير عن العلاقات البينية المرتبطة بخطط التصنيف من خلال استخدام نظام إحالات مكون من إحالة (انظر) و (انظر أيضاً) اللتين تُستخدمان عند الحاجة إليهما.

وقد تم استخدام خطط التصنيف كنموذج لتمثيل واسترجاع المعلومات الأحادية Monograph Information حيث يتم استخدام رمز تصنيف واحد للإشارة إلى كيان أو وعاء معلومات أو وحدة معلوماتية كاملة. ومن أبرز نماذج خطط التصنيف وأكثرها انتشاراً على المستوى العالمي كل من خطة تصنيف ديوي العشري وخطة تصنيف مكتبة الكونجرس واللتين تمت ترجمتهما إلى كل اللغات ومنها العربية.

- **خطوات التكشيف في نظم الربط المسبق**

تنطوي عملية التكشيف في نظم الربط المسبق على أربع مراحل أساسية هي:

1. التحليل المفاهيمي.
2. اختيار المصطلحات من لغة التكشيف المقيدة.
3. تركيب أو ربط المصطلحات معاً وفقاً لقواعد الربط التي توفرها لغة التكشيف .
4. إعداد الروابط التي تربط التسجيلة الببليوجرافية بمخزن الوثائق.

وناتج عملية التكشيف في هذه الحالة يتمثل في تراكيب مصطلحات مركبة أو معقدة، بالتالي ينبغي في عملية البحث أن تصاغ الرؤوس المستخدمة في البحث بنفس الطريقة التي أُعدت بها في أثناء عملية التكشيف لكي تتم عملية المضاهاة بين مصطلحات البحث والمصطلحات المستخدمة في عملية التكشيف. بمعنى آخر أنة ينبغي أن تكون الرؤوس أو الرموز المستخدمة في عملية البحث متطابقة تماماً مع الرؤوس أو الرموز المستخدمة في عملية التكشيف. وقد استخدمت هذه النظم في إعداد الفهارس الموضوعية الهجائية، الفهارس المصنفة، الببليوجرافيات الموضوعية المصنفة.

ومن أهم عيوب نظم تكشيف الربط المسبق ما يلي (لانكستر، 1997):

1. أنها معقدة من حيث البناء، حيث تتطلب إعداد تراكيب للمصطلحات تربط فيما بينها، بحيث ينتج في النهاية رأس موضوع واحد يعبر عن المحتوى الموضوعي لوعاء المعلومات.

2. هذا النوع من النظم يستخدم مدخلاً واحداً لترتيب المصطلحات المركبة أو المعقدة وهو ليس بالضرورة الرأس المناسب للبحث في كل الحالات. إضافة إلى أن وعاء المعلومات لا يمكن الوصول إليه إلا من خلال هذا

المدخل، بمعنى اختزال العلاقة بين المصطلحات في شكل خطي أو تتابع خطي باستخدام التوافيق المحتملة للمصطلحات، ما يقيد المستفيد في عملية البحث وفقاً لهذا التتابع الخطي.

3. أن هذا الأسلوب وإن كان اقتصادياً من حيث عدد المصطلحات المستخدمة في التعبير عن المحتوى الموضوعي للوثيقة، إلا أنه غير عملي، حيث إن زيادة عدد المصطلحات أو الفئات التي تنتمي إليها الوثيقة إلى 10 أو 15 مصطلحاً تخلق موقفاً يصبح من المستحيل فيه التعامل مع نظام الربط المسبق.

ومن الحلول التي طرحت للتغلب على مشكلات نظم الربط المسبق ما يلي:

- محاولات تشارلز كتر في استخدام مبدأ القلب في صياغة الرؤوس المركبة، الذي أوصى بوضع المصطلح الأهم في مقدمة الرأس، وذلك بقلب الرأس إذا لم يكن العنصر الأول فيه هو العنصر المهم. كما وضع أيضاً الجذور الأساسية لشبكة الإحالات التي تربط بين المصطلحات الواردة في لغة التكشيف كإحالات انظر وانظر أيضاً.

- استخدام فكرة التصنيف الوجهي: وتقوم فكرة التصنيف الوجهي على أساس أن كل الرؤوس المركبة أو المعقدة يمكن تركيبها باستخدام نسق عام لترتيبها يعتمد على تحديد العنصر المهم في الرأس، بحيث يأتي في البداية ثم يليه العنصر الأقل أهمية ثم الأقل أهمية. كما يرى كايزر أن رؤوس الموضوعات المركبة أو المعقدة يمكن تحليلها إلى مركب مكون من شيء محسوس Concrete، وعملية Process وأن المحسوس أو الشيء ينبغي دائماً أن يسبق العملية عند إعداد الرأس:

مثال الكتب – فهرسة

المكتبات – تنظيم

النظم – تحليل وتصميم

كما وضع مبادئ التفريعات الجغرافية والشكلية بحيث تلي تلك التفريعات العمليات التي تتم على المفهوم.

مثال: المكتبات – تنظيم – مصر (الشيء – العملية – التفريع الجغرافي)

المكتبات – مصر – أدلة (الشيء – التفريع الجغرافي – التفريع الشكلي)

كما قام رانجاناثان بإعداد أشهر خطة للتصنيف الوجهي في أواخر العشرينيات وأوائل الثلاثينيات من القرن السابق. وقد استندت فكرة رانجاناثان إلى تطوير أفكار كايزر للمحسوس والعملية، وذلك اشتمل على خمس فئات أساسية هي:

3.5 الشخصية: الشيء نفسه

3.6 المادة: مواد أساسية

3.7 الطاقة: عملية – أسلوب

3.8 المكان.

3.9 الزمان.

- **تدوير المصطلحات Term Rotation**

تستند فكرة تدوير المصطلحات إلى أساس إعطاء كل عنصر من عناصر الرأس فرصة الظهور في مقدمة الرأس. بالتالي يكون قابلاً للبحث والاسترجاع. وهي الفكرة التي استندت إليها فيما بعد كشافات الكلمات المفتاحية، فمثلاً إذا كان لدينا رأس موضوع معقد مثل نظم استرجاع المعلومات البيليوجرافية يمكن تدويره كاملاً باستخدام المعادلة التالية.

(N- I) X (N – 2) X (N- 3) X N

فإذا كان لدينا رأس مكون من أربعة مصطلحات، وعند تطبيق معادلة تدوير المصطلحات تكون كالتالي:

(4-1) X (4-2) X (4-3) X 4 = 3 X 2 X 4 = 24

أما إذا اشتمل الرأس على ثلاثة مصطلحات يكون عدد البدائل كما يلي:

(3-1) X (3-2) X 3 = 2 X 1 X 3 = 6

مثال: نظم استرجاع المعلومات

نظم استرجاع المعلومات

نظم المعلومات – استرجاع

استرجاع المعلومات – نظم

استرجاع – نظم – المعلومات

نظم المعلومات – استرجاع

نظم – استرجاع المعلومات

وتجدر الإشارة إلى أن من أهم عيوب عملية تدوير المصطلحات الزيادة الكبيرة في عدد البدائل، ما يؤدي إلى تضخم الكشافات، مع العلم أن تلك الآليات كانت تستخدم مع الكشافات المطبوعة للتغلب على مشكلات اللغة المضبوطة.

5.1.3.2 نظم تكشيف الربط اللاحق

Post Coordinate Indexing Systems

هي النظم التي يتم الربط فيها بين المصطلحات التي تمثل المفاهيم المختلفة لكي تظهر في شكل رؤوس مركبة أثناء عملية البحث والاسترجاع. في هذه النظم يتم تمثيل المصطلحات التي تعبر عن المفاهيم الواردة في الوحدة المكشفة في صورة مصطلحات مفردة، فيما يطلق عليه نظام المصطلح الواحد Uniterm دون الحاجة إلى إعداد تراكيب مصطلحات معقدة أثناء عملية التكشيف. ما يقضي على مشكلة التتابع الخطي للمصطلحات، ويقضي بالتبعية على الحاجة إلى تدوير المصطلحات. كما أنه يوفر إمكانية الوصول إلى الوثائق باستخدام المصطلحات المفردة والمصطلحات المركبة والمعقدة.

وقد اتخذت نظم الربط اللاحق أشكالاً متعددة في مراحلها الأولى؛ منها الاعتماد على البطاقات المثقبة في تمثيل المصطلحات المفردة، ما أدى إلى ظهور مبادئ مختلفة للمضاهاة أو المطابقة بين المصطلحات المستخدمة في عملية التكشيف والمصطلحات المستخدمة في عملية البحث والاسترجاع. ومن هذه الأساليب مبدأ المطابقة البصرية ومنها أيضاً مبدأ المطابقة الميكانيكية. وقد اعتمد كل منهما على استخدام بطاقة واحدة للتعبير عن المصطلحات المختلفة في النظام فيما يعرف ببطاقة الوثيقة أو استخدام بطاقة واحدة لكل مصطلح فيما يعرف ببطاقة المصطلح.

وتجدر الإشارة إلى أن طرق إعداد بطاقة المصطلح وبطاقة الوثيقة قام بتطويرها كل من باتن Batten ومورز Mooers في نهاية الأربعينيات من القرن الماضي، ولم تزل الطريقتان هما الأساسيتين في بناء ملفات النظم الإلكترونية المعتمدة على الحاسبات الآلية في استرجاع المعلومات.

ومن أهم الملامح العامة التي تتميز بها نظم التكشيف الربط اللاحق أنها:ـ

1. تعالج موضوعات الوثائق كمفاهيم فردية يتم التعبير عنها باستخدام نظام المصطلح الواحد دون الحاجة إلى توافيق أو تراكيب مصطلحات معقدة.
2. تعتمد هذه النظم على اختيار المصطلحات من لغة تكشيف مضبوطة أو مقننة يُطلق عليها المكانز سوف نتناولها بالتفصيل فيما بعد.
3. يجب استخدام لغة التكشيف المضبوطة أيضاً لاختيار المصطلحات المناسبة للتعبير عن المفاهيم الواردة في استفسارات المستفيدين.
4. بعد اختيار مصطلحات البحث من لغة التكشيف المضبوطة يتم الربط بينها في مرحلة البحث والاسترجاع من خلال إعداد استراتيجية البحث.
5. هذه الطريقة تمثل الأساس الذي تعتمد عليه معظم النظم الإلكترونية في تمثيل الوثائق، بالتالي فهي تصلح أساساً لنظم استرجاع المعلومات المعتمدة على الحاسبات الآلية.

6. تتمتع هذه النظم بالمرونة الكافية، حيث إنه يمكن تمثيل محتويات الوحدة المكشفة بأي عدد من المصطلحات، بالتالي يمكن تحقيق مستوى العمق اللازم عند تكشيف الوثائق دون الحاجة إلى إعداد توافيق مركبة أو معقدة معتمدة على التتابع الخطي للمصطلحات، كذلك دون الحاجة إلى تدوير المصطلحات من أجل تيسير عملية الوصول إليها.

- **المكانز**

المكنز عبارة عن قائمة مصطلحات مضبوطة تعتمد في صياغتها للمصطلحات على أسلوب المصطلح المفرد Uniterm القائم بذاته، بحيث يمكن ربطه بغيره من المصطلحات عن طريق معاملات البحث فيما يطلق عليه الربط اللاحق (Post – coordination National Information Standards Organization,1993)، كما عرفها رولي بأنها: قائمة بالمصطلحات والعبارات توضح المترادفات والبناء الشجري وغيرهما من العلاقات ومدى تبعية مصطلح لمصطلح آخر، والتي تساعد على توفير قائمة معيارية لخزن واسترجاع المعلومات (Rowley,1992, P.25).

ويعد الربط اللاحق أحد آليات معالجة المصطلحات في نظم استرجاع المعلومات التي ظهرت كبديل لنظم الربط المسبق التي تعتمد على خطط التصنيف وقوائم رؤوس الموضوعات. وتساعد نظم الربط اللاحق المستفيدين على إقامه علاقات بين المصطلحات وإنشاء تراكيب البحث في مرحلة تمثيل واسترجاع المعلومات. ومن أبرز عيوب الربط اللاحق هو الربط الخاطئ، وأحد أبرز الأمثلة على ذلك مصطلحان مثل Desk , Computer يمكن ربطهما بطريقتين مثل Computer Desk أو Desk Computer وذلك بناء على الغرض الأساسي من الموضوع، فإذا كان الباحث يريد معلومات عن Desk Computer فإن النتائج التي يكون الربط فيها Computer Desk سوف تؤدي إلى ربط خاطئ ونتائج غير دقيقة.

وتستخدم الحواشي المعيارية في بناء المكانز لتحديد العلاقات الشجرية (الهرمية) وعلاقات الارتباط وغيرها من العلاقات بين المصطلحات. وتستخدم شبكات الإحالات لتحديد المصطلحات المفضلة في الاستخدام للدلالة على الموضوعات

والمفاهيم مثل إحالة مستخدم Use وإحالة مستخدم لـ Used for (UF). وتستخدم حواشي المجال Scope Note (SN) في تحديد نطاق استخدام المصطلح والمعنى الدلالي للمصطلح المستخدم. ويتم توضيح العلاقات الشجرية بين المصطلحات من خلال علاقات البناء الهرمي للمصطلح الأضيق Narrower Term (NT)، المصطلح الأوسع Broader Term (BT)، كما يتم التعبير عن علاقات الارتباط Associative Relationship من خلال استخدام إحالة المصطلح المرتبط Related Term (RT).

وعادة ما يتم ترتيب المكانز ترتيباً هجائياً وهرمياً لتيسير الوصول إلى شبكة المصطلحات وعلاقاتها ببعضها بعضاً. كما يتم أحياناً استخدام أساليب التدوير Rotated والتبديل Permuted في عرض المصطلحات إلى جانب أساليب العرض النظمي Systematic أو التصنيفي Classification أو العرض الشكلي Graphical لاستعراض المصطلحات وعلاقاتها ببعضها بعضاً. (Aitchison, Gilchrist & Bawden , 1997).

وتعد المكانز أكثر قوائم اللغات المضبوطة شيوعاً في الاستخدام في نظم تمثيل واسترجاع المعلومات، حيث تعتمد قوائم رؤوس الموضوعات وخطط التصنيف التحليلية التركيبية على نظم المصطلح غير الواحد Non-Monograph، ما يحد من مرونة تلك الأنظمة، بينما تتميز المكانز التي تعتمد على نظم المصطلح الواحد بالمرونة إلى جانب قدرتها على معالجة المفاهيم المعقدة، من خلال معاملات الربط والعلاقات المتنوعة والإحالات.

وقد اتجهت منذ بداية القرن الواحد والعشرين العديد من قوائم رؤوس الموضوعات نحو التحول إلى استخدام البناء المكنزي في التعامل مع المصطلحات وشبكة الإحالات، ولعل أبرز مثال على ذلك قائمة رؤوس موضوعات مكتبة الكونجرس وقائمة رؤوس الموضوعات الطبية[1].

---

(1) https://www.nlm.nih.gov/mesh/filelist.html

5.1.4 مقارنة بين المكانز وقوائم رؤوس الموضوعات وخطط التصنيف

يشير جدول (4.1) إلى ملخص للملامح المميزة للأنواع الثلاثة المستخدمة في تمثيل نظم اللغة المضبوطة. فإلى جانب ما تم مناقشته لاحقاً، فإن لغات الربط المسبق تتميز بملمح مهم آخر يتمثل في طرق التحليل. ولعل أبرز طرق التحليل التي تتبعها تلك الأدوات أنها أدوات حصر Enumeration Tools، ، ما يعني أنها تتيح قوائم حصرية بالمصطلحات التي تمثل الإطار المعرفي الكامل سواء كان طبيعياً (كما هو الحال في المكانز وقوائم رؤوس الموضوعات) أو مصطنعاً كما هو الحال في خطط التصنيف دون الحاجة إلى دمج المصطلحات معاً للتعبير عن إطار معقد للمعرفة. وعلى العكس من ذلك، تعد تلك اللغات أيضاً أدوات تركيب Synthesis Tools تتيح الدمج بين المصطلحات لبناء تركيب أكثر تعقيداً سواء كان ذلك في مرحلة التمثيل أو البحث (لانكستر، 1997) ويوجد ارتباط جذري بين طريقة التحليل وطريقة الربط في تلك الأدوات. ويرجع ذلك إلى أن أدوات الربط المسبق تُعد أدوات حصرية في بنيتها، بينما تُعد أدوات الربط اللاحق أدوات تحليلية تركيبية. ويتم تحديد مستويات التحليل والربط ومدى التخصيص والمرونة في قوائم الربط المسبق من خلال مبادئ للربط وإقامة العلاقات، بينما تتميز أدوات الربط اللاحق بوجود مرونة في آليات الربط وعدم الثبات في مستويات التخصيص، ما يجعلها أكثر تخصيصاً واستيعاباً للجوانب المعرفية المتنوعة من لغات الربط المسبق الحصرية. ووفقاً للجدول (4.1) فإن المكانز تُعد أكثر اللغات المضبوطة تخصيصاً ومرونة في الاستخدام من كل من خطط التصنيف وقوائم رؤوس الموضوعات؛ ما يفسر لماذا تُعد المكانز أكثر لغات المصطلحات المضبوطة انتشاراً واستخداماً في تمثيل واسترجاع المعلومات.

جدول 4.1 مقارنة لغات المصطلحات المضبوطة

اللغة / الخاصية	المكانز	قوائم رؤوس الموضوعات	خطط التصنيف
مكونات المصطلح	واصفات	رؤوس موضوعات	رموز التصنيف
أسلوب الإحالات والحواشي	استخدم، مستخدم لـ	استخدم، مستخدم لـ مستخدم بدلاً من، مستخدم بدلاً من أيضاً	انظر وانظر أيضاً
طرق التحليل	تحليلية تركيبية		حصرية
طرق الربط	لاحق	مسبق ولاحق	مسبق
التخصص	أكثر تخصصياً	مخصصة إلى حد ما	عامة
المرونة	أكثر مرونة	مرنة إلى حد ما	أقل مرونة
المواد المستهدفة	المنفردات والمواد التحليلية	المنفردات المواد التحليلية	المنفردات

5.2 نظم تكشيف اللغة الطبيعية

تعمل نظم التكشيف بصفة عامة على إعداد بدائل للوثائق يمكن بحثها بسهولة من خلال المقارنة أو المطابقة بين المصطلحات الواردة في استفسارات المستفيدين والمصطلحات التي تم اختيارها للتعبير عن المحتوى الموضوعي للوثائق. فإذا كانت نظم التكشيف المضبوطة أو المقيدة تنتقي مصطلحات التكشيف من أدوات أو لغات تكشيف معدة ومجهزة مسبقاً، فإن نظم تكشيف اللغة الطبيعية تنتقي المصطلحات التي تستخدم للتعبير عن الوحدات المكشفة مباشرة من النصوص التي يتم تكشيفها دون الاعتماد على أدوات مقيدة لضبط المصطلحات والتحكم فيها، سواء تم هذا الاختيار يدوياً من قبل المكشف أو آلياً من خلال برنامج للحاسب الإلكتروني.

تستند هذه النظم إلى مبدأ أساسي هو أن مؤلفي الوثائق عادة ما يستخدمون

مصطلحات محددة للتعبير عن الأفكار التي يريدون توصليها. وهذه المصطلحات عادة ما تكون شائعة ومعروفة في المجالات التي يعملون بها. وينطبق هذا المبدأ بشكل أكثر دقة على المجالات العلمية والتكنولوجية، بمعنى أن المؤلفين عادة ما يتواصلون مع مجتمع القراء من خلال لغة شائعة ومعروفة لجميع المتخصصين في هذه المجالات. بالتالي يكون إقحام لغة وسيطة (اللغة المضبوطة) في هذه العملية أمراً اصطناعياً ينتج عنه وجود حاجز بين المؤلف والقارئ يتمثل في تلك اللغة الاصطناعية.

فبالنظر إلى عملية التكشيف اليدوية التي تعتمد على الجهد البشري نجد أنه من الممكن التعرف إلى المفاهيم التي تتناولها الوثائق من خلال التحليل المفاهيمي للمحتوى المحوري في الوثيقة، والذي يظهر في مواضع محددة مثل العناوين وقوائم المحتويات والمستخلص ورؤوس الموضوعات الجانبية ومقدمة النص.. الخ. ومن خلال فحص تلك المواضع وتحديد الأهمية النسبية (التي عادة ما تستخدم فيها معايير كمية وكيفية، مثل تردد المصطلح وأهمية المصطلح للمستفيدين وعلاقته بدور المؤسسة)، لكل مفهوم ورد في تلك المواضع يحدد المكشف المصطلحات التي تستخدم في تكشيف الوثيقة. وعلى افتراض أن النص متاح في شكل إلكتروني، بالتالي يكون من السهل إعداد برمجيات مصممة خصيصاً لكي تقوم بالتكشيف الاشتقاقي من خلال الاعتماد على المبادئ السابقة نفسها مثل تردد المصطلحات Term Frequency، موضع المصطلح Term Position، وغيرها من المعايير التي يمكن الاعتماد عليها في بناء خوارزميات تحدد أهمية المصطلح بالنسبة للوثيقة التي يتم تكشيفها.

ويمكن تتبع بداية نظم التكشيف الآلي المعتمدة على مبدأ تردد المصطلحات إلى الخمسينيات من القرن العشرين وخاصة أعمال لوهان وباكسندال. فقد شهدت تلك الفترة بدايات الاعتماد على الحاسب الإلكتروني في إعداد النصوص للنشر. من هنا بدأت فكرة استخدام الحاسب الآلي في عمليات البحث والاسترجاع في الظهور، حيث وجد أنه مادامت النصوص متاحة أصلاً في شكل إلكتروني، يمكن الاعتماد على هذه النصوص الإلكترونية في عمليات التكشيف والاستخلاص والاسترجاع. من ثم فإن التطورات في مجال الحاسبات الآلية ساعدت بشكل كبير على تكشيف النصوص آلياً

بالاعتماد على اشتقاق المصطلحات من اللغة الطبيعية التي يستخدمها المؤلفون في التعبير عن أفكارهم بشكل أكثر سهولة وسرعة. كما أنه أقل في الكلفة من نظم التكشيف اليدوية، ما يحقق فعالية وعائداً من خدمات التكشيف والاستخلاص (Luhn, 1958).

وقد ساعد على تطوير نظم اللغة الطبيعية عاملان أساسيان هما:

1. التطوير المذهل في تقنيات الحاسب الآلي التي ساعدت على تخزين النصوص الكاملة للكتب والدوريات وغيرها من أوعية المعلومات حتى أصبح مجال النشر الإلكتروني هو النمط السائد عالمياً في النشر والتوزيع، ما ساعد على تيسير معالجة النصوص من حيث حجم الاختزان وسرعة المعالجة.
2. التطور المذهل في مجال البرمجيات، والذي ساعد على إعداد برامج مصممة خصيصاً لكي تقوم بعمليات التكشيف الآلي، ولا شك أن هناك نظم استرجاع معلومات تستطيع الآن معالجة النصوص باللغة الطبيعية بدرجة عالية من الدقة والكفاءة.

وقد ساعد استخدام نظم اللغة الطبيعية في عمليات التكشيف على التخلص من عمليات البحث المفوض الذي يقوم فيه وسيط بين نظام الاسترجاع والمستفيد بعمليات البحث والاسترجاع، حيث أصبحت معظم نظم استرجاع المعلومات الآن تتضمن واجهات تعامل صديقة للمستفيد يمكن من خلالها التفاعل بين المستفيد والنظام دون الحاجة إلى وسيط يساعد على إعداد الاستفسارات وبناء استراتيجيات البحث وإجراء البحث نيابة عن المستفيدين.

إذاً، فاللغة الطبيعية هي اللغة التي يستخدمها البشر في الحديث والكتابة، وعند تطبيقها في نظم استرجاع المعلومات يتم اشتقاق المصطلحات من الوثائق للتعبير عن المفاهيم ومضمون ومحتوى الوثائق. وتعتمد عملية الاشتقاق على أساليب رياضية أو إحصائية لتحديد أهم المصطلحات المستخدمة بالوثائق للدلالة على المفاهيم. ولا تحتاج نظم تمثيل واسترجاع المعلومات إلى بذل مجهود لتحديد أو تعريف المصطلحات سواء من الناحية البنائية Syntax أو الدلالية Semantic أو

العلاقـات المتداخلـة Interrelationships بيـن المصطلحـات. فاللغـة الطبيعيـة تشـير إلـى مـا يسـتخدمه النـاس فـي التعبيـر عـن المعلومـات أو صياغـة الاسـتفسـارات دون الرجـوع إلـى لغـة مضبوطـة لتقنيـن المصطلحـات.

5.2.1 طرق التمثيل باللغة الطبيعية

وتوجـد ثـلاث طـرق أساسـية لاسـتخدام اللغـة الطبيعيـة بصفـة عامـة لأغـراض تمثيـل واسـترجاع المعلومـات هـي كالتالـي:

5.2.1.1 اشتقاق الأجزاء

تعتمـد هـذه الطريقـة علـى تحديـد أهـم المصطلحـات الـواردة فـي الوثيقـة واشـتقاقها مـن أبـرز الأجـزاء التـي تمثـل المحتـوى أو التـي يركـز عليهـا منشـئ الوثيقـة. وتعـد العناويـن أهـم أجـزاء الوثائـق، لذلـك يتـم توظيفهـا فـي تحديـد أهـم المصطلحـات التـي تعبـر عـن محتـوى الوثائـق. وقـد اسـتخدمت العناويـن فـي تمثيـل محتـوى الوثائـق مـن خـلال بنـاء كشـافات العناويـن، والتـي ابتكرهـا لوهـان هانـز بيتـر Luhn Hans Peter فـي بدايـة السـتينات مـن القـرن الماضـي. وقـام بتطبيقهـا علـى البطاقـات المثقبـة باسـتخدام آليـات المضاهـاة الضوئيـة والميكانيكيـة فـي مكتبـات مانشسـتر فـي عـام 1864. وتعـد كشـافات العناويـن نموذجـاً فريـداً لمـا يطلـق عليـه كشـافات التباديـل Premuted Index. ويشـير المصطلـح إلـى تطبيـق مفهـوم التدويـر ومبـدأ التباديـل الدائريـة cyclic permutations للـرؤوس، مـا يتيـح للمسـتفيد البحـث عـن أي كلمـة مـن الكلمـات الـواردة فـي الـرأس. وقـد تـم تطبيـق هـذا المبـدأ علـى عناويـن الوثائـق، ونتـج عـن هـذا الأسـلوب ثـلاث طـرق لتكشـيف العناويـن، سـيتم شـرحها بالتفصيـل عنـد تنـاول طـرق عمـل نظـم تكشـيف اللغـة الطبيعيـة وهـي:

- كشاف الكلمات المفتاحية في السياق Key Words In Context (KWIC)
- كشاف الكلمات المفتاحية خارج السياق Key Words Out of Context (KWOC)
- كشاف الكلمات المفتاحية المضافة للسياق Key Words Added to Context (KWAC)

كما يستخدم مع اشتقاق عبارات الموضوع Topic Sentence أو غيرها من الأجزاء المهمة التي تأتي في صورة عبارات وجمل يمكن أن تستخدم في تمثيل الوثيقة (Luhn, Hans Peter, 1960).

5.2.1.2 اشتقاق المصطلحات

تعتمد تلك الطريقة على اشتقاق كلمات من أي جزء من أجزاء النص فيما يطلق عليه التكشيف الاشتقاقي Indexing Derivative. وعادة ما يتم تطبيق خوارزميات متنوعة لتحديد أهم المصطلحات الدالة على المفاهيم التي تناولتها الوثيقة. ولعل أبرز هذه الخوارزميات ما يلي:

- تردد المصطلحات Term Frequency
- مواضع المصطلحات Term Position
- تردد المصطلح في الموضع Term Frequency Vs. Psition
- الوزن ن ـ N gram
- وزن المصطلح Term Wight

وتستخدم كل هذه الأساليب الإحصائية في تحديد أهم المصطلحات الدالة على المفاهيم التي تعالجها الوثيقة، بالاعتماد على فرضية أساسية هي: أنه كلما ارتفعت معدلات تردد مصطلح معين في وثيقة معينة، فإن هذا يعد مؤشراً أساسياً على أهمية هذا المصطلح في هذه الوثيقة.

5.2.1.3 اشتقاق الأسئلة

يستخدم هذا الأسلوب في نظم الرد على الاستفسارات، ويعتمد هذا النموذج على الكلمات والعبارات المشتقة مباشرة من أسئلة البشر المستخدمة في تمثيل الاستفسارات Query Representation.

وتتكون اللغة الطبيعية بصفة عامة من نوعين من الكلمات هما:

- الكلمات الفريدة Significant words
- الكلمات الوظيفية Function words

الكلمات الفريدة هي الكلمات التي تستخدم كمصطلحات تحمل معاني ودلالات موضوعية، أما الكلمات الوظيفية فهي الكلمات التي تشير إلى حروف الجر، التذكير والتأنيث، حروف الوصل، أدوات التعريف والتنكير Articles, Proposition Conjunction مثل في اللغة الإنجليزية an, a, the, and, for, of, to, this, that, her, their. ويتم توظيف هذين النوعين من الكلمات في عمليات التمثيل من خلال اشتقاق الكلمات الفريدة ووضعها في كشاف، واستبعاد الكلمات الوظيفية ووضعها في قائمة استبعاد Stop – Word – List أو Stop List.

وتستخدم قوائم الكلمات الفريدة في تحديد الكلمات التي يتم تكشيفها ومصطلحات الاستفسار، والتي عادة ما يتم التعبير عنها بأنها أي كلمة لم ترد في قائمة الاستبعاد. وتتضمن قوائم الاستبعاد الكلمات الوظيفية كثيرة التواتر إضافة إلى أي كلمة فريدة عامة كثيرة التواتر في مجال ما أو شائعة الانتشار في لغة البشر. فعلى سبيل المثال مصطلح Engineering يعد مصطلحاً عاماً في أي قاعدة بيانات هندسية إلى جانب الكلمات ذات الطبيعة العابرة Ephemeral words مثل الكلمات الطنانة Buzz words مثل من ثم، مما لا شك فيه، على سبيل المثال، هذه الكلمات أيضاً يتم وضعها في قائمة الاستبعاد ولا يتم توظيفها في عملية التكشيف والاسترجاع.

ويقوم كل نظام تمثيل واسترجاع معلومات ببناء قائمة الاستبعاد الخاصة به بناء على احتياجات المستفيدين منه وطبيعة المواد المكشفة بالنظام. كما يتم بناء قائمة مناظرة لقائمة الاستبعاد يُطلق عليها قائمة الذهاب Go List. وتشتمل تلك القائمة على كل المصطلحات الواردة في الوثيقة بعد استبعاد الكلمات الواردة في قائمة الاستبعاد والعبارات الطنانة كثيرة التواتر (Rowley, 1992).

وكما هو الحال في قائمة الاستبعاد فإن قائمة الذهاب يتم تجميعها وقراءتها

آلياً، كما يتم مقارنتها بكل وثيقة يتم تمثيلها واستفسار يتم بحثه. ومن المعروف أن هذه القوائم تنمو بصفة دائمة مع نمو نظام استرجاع المعلومات. ومع ذلك فإن قوائم الذهاب أقل استخداماً في نظم اللغة الطبيعية من قوائم الاستبعاد التي تعد أكثر انتشاراً نظراً لسهولة إعدادها ووجود نماذج عامة لها إلى جانب انخفاض كلفة بنائها مقارنة بقوائم الذهاب. من ثم فإن قوائم الاستبعاد تتميز بأنها:

- أقل في الحجم من قوائم الذهاب
- سهوله إدارتها (التجميع والمعالجة)
- قوائم الذهاب تستخدم في بناء لغات التكشيف المضبوطة مثل المكانز وقوائم رؤوس الموضوعات وخطط التصنيف.

وفي السنوات الأخيرة بدأت بعض النظم بناء قوائم كلمات Word lists وهي قوائم مصطلحات شبه مضبوطة Semi Controlled Vocabulary في النظم الآلية لتمثيل واسترجاع المعلومات. وتشتمل قوائم الكلمات على المترادفات Synonyms والمتضادات Antonyms للمصطلحات الواردة في الوثائق التي يتم تكشيفها ويتم توظيفها في دعم المستفيد أثناء عمليات البحث والاسترجاع. وتعد هذه النوعية من القوائم نموذجاً فريداً لقوائم الذهاب التي تستخدم في ضبط عمليات البحث للتغلب على مشكلات الترادف والاشتراك اللفظي والبحث الشامل التي تواجهها نظم اللغة الطبيعية.

وتقوم العديد من نظم استرجاع المعلومات على الإنترنت مثل محركات بحث الويب ببناء قوائم ذهاب وقوائم كلمات لاستخدامها في ضبط المصطلحات وضبط دلالتها. فمع النمو الهائل للويكيبيديا، أصبح من الممكن اعتماد قوائم مصطلحاتها كنموذج أساسي لقوائم الكلمات التي يمكن أن تكون أكثر كفاءة من أي أداة أخرى.

5.2.2 أسلوب عمل نظم تكشيف اللغة الطبيعية

تعتمد تلك النظم ببساطة على أنظمة التكشيف الآلية التي تقوم بإحصاء عدد مرات تردد المصطلحات في النص من خلال اتباع الخطوات التالية:

1. إعداد ملف بالكلمات المستبعدة Stop List يشتمل على الكلمات كثيرة التواتر More Frequently Repeated Terms في النصوص والتي لا تحمل دلالة اصطلاحية مثل حروف الجر أدوات التعريف والتذكير والتأنيث وغيرها والتي سبق ذكرها.

2. يقوم نظام التكشيف الآلي بقراءة كلمات النص أولاً لاستبعاد الكلمات التي تتطابق مع الكلمات الواردة في قائمة الاستبعاد.

يساعد استخدام قوائم الاستبعاد على تحقيق ما يلي:

- تصغير حجم الكشاف.
- سرعة عملية التكشيف.
- الفعالية، حيث لا يتضمن الكشاف إلا الكلمات القابلة للبحث.

ومن الجدير بالذكر أنه عند تكشيف أنواع معينة من النصوص التي يكون لكل كلمة فيها أهمية ودلالة معرفية مثل النصوص الدينية، التشريعات، المعادلات الكيمائية والرياضية.. الخ، لا يتم استخدام قوائم الاستبعاد أثناء عمليات التكشيف.

3. يقوم نظام التكشيف الآلي بحساب عدد مرات تردد كل مصطلح في الوثيقة، ثم ترتيب تلك المصطلحات وفقاً لعدد مرات ورودها في النص، بحيث ترد المصطلحات الأكثر تردداً على قمة القائمة تليها المصطلحات الأقل فالأقل.

4. يتم اختيار مجموعة محددة من المصطلحات وفقاً لنقطة القطع Cutoff Point المحددة بالنظام. وهي النقطة التي تحدد عدد المصطلحات التي يتم اختيارها، ويمكن أن تعتمد تلك النقطة على مجموعة من المعايير أو الاحتمالات منها:

- رقم مطلق لعدد المصطلحات مثال اختيار أكثر 20 مصطلحاً تردد في الوثيقة.
- رقم مرتبط بطول الوثيقة بحيث يكون عدد المصطلحات الوثائق الكبيرة في الحجم أكبر من عدد مصطلحات الوثائق الأقل حجماً. مثال وثيقة حجمها

4000 كلمة نختار أعلى 20 مصطلحاً، أما إذا كان حجم الوثيقة 2000 كلمة فيتم اختيار أعلى 10 مصطلحات لوصفها.

- اختيار المصطلحات التي وردت في أماكن محددة من الوثيقة و/ أو عدد مرات ورودها في تلك الأماكن.

5. يمكن لبعض البرامج الأكثر تعقيداً أن تنتقي أو تشتق العبارات التي تظهر بشكل متكرر في بعض النصوص. لذلك يمكن وصف الوثائق باستخدام مزيج من المصطلحات والعبارات. وتجدر الإشارة إلى أن عدد مرات ظهور العبارة يكون أقل أهمية من عدد مرات ظهور المصطلح. وبدلاً من اختيار المصطلحات والعبارات يمكن لبعض البرامج أن تقوم بتجريد الكلمات واختيار جذور تلك الكلمات فقط Word Roots وذلك بالاعتماد على برنامج للجذع يعرف بـ Stemmer. لذلك فإن جذر الكلمة Heat يمكن أن يشتق ويخزن لكل بدائل هذه الكلمة التي تشمل Heat, Heater, Heating, Heated بالتالي فإن برامج الجذع الآلي تستخدم لحذف نهايات وبدايات الكلمات Word Suffix and Prefix مثل ing, ed, ied, pre, sub, s, es, ies. وفي اللغة الغربية نجد أنه يمكن جذع بدايات ونهايات الكلمات مثل الألف واللام، الألف والنون (للمثنى) الياء والنون والألف والنون للجمع إلى آخره من المتطلبات التي تفرضها طبيعة وبنية الكلمات في اللغة العربية.

6. يمكن إعطاء الكلمات أو الجمل أو جذوع الكلمات وزناً معيناً يعكس عدد مرات تردد المصطلح في الوثيقة. على سبيل المثال يمكن إعطاء الجذع Heat وزناً معيناً يحدد أنه ظهر في نص معين 12 مرة. وتصلح عملية جذع الكلمات بشكل أكبر للغات اللاتينية، حيث توصف بأنها لغات لصيقة غروية. بمعنى أنها تستخدم أسلوباً محدد لاشتقاق الكلمات بإضافة حروف معينة في بداية الجذر أو نهايته في معظم الأحوال، بينما يلاحظ أن اللغة العربية لا تخضع لهذا النموذج اللصقي في بناء الكلمات، حيث تعرف بأنها لغة اشتقاقية نظراً لتنوع الصيغ الخاصة بمعالجة مفردات اللغة مثل الفعل والفاعل والمفعول،

حيث تعتمد اللغة العربية على قواعد متنوعة ومتشعبة بصورة كبيرة تميل إلى السماع أكثر منها إلى الثبات في البنية في معالجة المفردات، كما هو الحال في معظم مفردات اللغات اللاتينية.

5.2.3 أنماط نظم تكشيف اللغة الطبيعية

توجد أنماط عدة لنظم تكشيف اللغة الطبيعية ولكن أشهرها وأكثرها انتشاراً على الإطلاق الأنماط التالية:

1. كشافات أو فهارس النصوص Concordances
2. كشافات العناوين التبادلية Permuted Title Indexes
3. التكشيف الآلي Automatic Indexing

وسنتناول فيما يلي بإيجاز هذه الأنماط المختلفة.

5.2.3.1 كشافات النصوص

تعد كشافات النصوص للوثائق التي تتضمن نصوصاً مهمة مثل النصوص الدينية، والتي يكون لكل كلمة في النص قيمتها، بحيث لا يمكن استبعادها من عمليات التكشيف. بالتالي فهذه الكشافات لا تستخدم قوائم استبعاد، حيث يتم تكشيف كل كلمات النص دون تمييز بينها. كما تستخدم هذه الكشافات أيضاً مع النصوص الصغيرة مثل الدساتير والتشريعات والقرارات والوصفات.. إلخ.

ويتطلب إعداد كشافات النصوص أن يكون النص المُكشف مُتاحاً في شكل مقروء آلياً. وقد ساعد النشر الإلكتروني على توافر عدد كبير من النصوص في صيغ رقمية، ما ييسر عمليات تكشيف نصوصها. ويتيح هذا النوع من الكشافات الوصول إلى المعلومات الدقيقة المتضمنة في النصوص الكاملة للوثائق وليس مجرد إشارات ببليوجرافية إلى الوثائق. كما ييسر هذا النوع من الكشافات عمليات التحليل اللغوي للنصوص للتعرف إلى تردد الكلمات والمصطلحات في سياقات معينة بهدف تحديد الدلالات المختلفة.

ويعد «المعجم المفهرس لألفاظ القرآن الكريم» لمحمد فؤاد عبد الباقي. و«المورد المفهرس لألفاظ القرآن الكريم» لروحي البعلبكي، من أشهر أنواع كشافات النصوص في اللغة العربية. وتجمع هذا المعاجم ألفاظ القرآن، وترتب موادها، كما تضع الكلمة وأمامها الآية الكريمة التي وردت فيها، مع التنبيه على المكي والمدني من هذه الآيات وحسب ما ورد في المصحف، الذي تولت الحكومة المصرية طبعه. وقد رتب عبد الباقي جميع ألفاظ القرآن الكريم ترتيباً هجائياً حسب مواد الكلمات الدالة، ثم سرد الألفاظ، وذكر تحت كل لفظة عدد مرات ورودها في القرآن حسب الصيغة الإعرابية والاشتقاقية التي وردت بها. فإذا وردت الكلمة بصيغة واحدة فإنه يترك الإشارة إلى عدد مرات ورودها (حسام الدين، 1994).

وتجدر الإشارة إلى أن المستشرق جوستاف فلوجل، هو أول من حاول إعداد معجم مفهرس لألفاظ القرآن الكريم، حيث قام بإصدار فهرسٍ موضوعيٍّ لآيات القرآن الكريم سمّاه «نجوم الفرقان في أطراف القرآن» في نحو عام 1868 (عام 1257هـ) ـ وقصد من وراء هذا المعجم ـ بحسب رأي بعض الباحثين ـ إعادة ترتيب القرآن حسب الموضوعات، وقد مهّد لمشروعه في تأليف معجمه الموضوعيّ «نجوم الفرقان في أطراف القرآن» بطباعة مصحفٍ كامل لكي يستعين به في معجمه، فوقع في أخطاء فاحشة وكثيرة جداً في عدّ الآيات، فجعل ما ليس برأس آيةٍ رأسَ آية، ووقع الخلل في معجمه بشكل ظاهر (جلغوم، 2012).

ومن أهم عيوب كشافات النصوص، خصوصاً اليدوية منها، أنها تحتاج إلى وقت وجهد كبيرين لإنجازها، إضافة إلى صعوبة بنائها وتضخم حجمها، حيث يتعدى حجمها في أحيان كثيرة حجم النصوص الأصلية.

5.2.3.2 كشافات العناوين التبادلية

يعتمد هذا النوع من كشافات اللغة الطبيعية على تكشيف كلمات العناوين بعد استبعاد الكلمات الواردة في قائمة الاستبعاد. وتستند كشافات العناوين إلى فكرة أساسية مفادها أن عناوين الوثائق تحتوي على كلمات أو مصطلحات تدل بشكل دقيق

على المحتوى الموضوعي للوثيقة وخصوصاً في المجالات العلمية والتكنولوجية. بالتالي يمكن استخدام هذه المصطلحات في وصف المحتوى الموضوعي الوثائقي. ولهذا النوع من الكشافات ثلاث أنماط أساسية كما أشرنا هي:

- كشافات الكلمات الدالة في السياق keyword In Context (KWIC).
- كشافات الكلمات الدالة خارج السياق keyword Out Of Context (KWOC).
- كشافات الكلمات الدالة المضافة للسياق keyword Add to Context (KWAC).

أ. كشافات الكلمات الدالة في السياق

يتم تكشيف الكلمات الدالة في عناوين الوثائق، حيث ترد الكلمة ضمن سياق العنوان مميزة عن غيرها من الكلمات.

مثال مقالة بعنوان

«استخدام الحاسب الآلي في تطبيقات المكتبات» وأخرى بعنوان

«تطبيقات تكنولوجيا المعلومات في المكتبات»

يشتمل كلا العنوانين السابقين على كلمة واحدة يمكن أن ترد بقائمة الاستبعاد هي (في) بالتالي يكون شكل الكشاف كما يلي:

استخدام الحاسب الآلي في تطبيقات المكتبات	(1)
استخدام الحاسب الآلي في تطبيقات المكتبات	(1)
تطبيقات تكنولوجيا المعلومات في المكتبات	(2)
تطبيقات تكنولوجيا المعلومات في المكتبات	(2)
استخدام الحاسب الآلي في تطبيقات المكتبات	(1)
تطبيقات تكنولوجيا المعلومات في المكتبات	(2)
استخدام الحاسب الآلي في تطبيقات المكتبات	(1)
تطبيقات تكنولوجيا المعلومات في المكتبات	(2)

ب. كشافات الكلمات الدالة خارج السياق

ترد الكلمات الدالة في هذا الشكل خارج السياق مميزة عن بقية العنوان مثال:

استخدام الحاسب ؟ في تطبيقات المكتبات	الآلي
الحاسب الآلي في تطبيقات المكتبات؟	استخدام
تكنولوجيا المعلومات في المكتبات؟	تطبيقات
استخدام الحاسب الآلي في ؟ المكتبات	تطبيقات
تطبيقات؟ المعلومات في المكتبات	تكنولوجيا
استخدام؟ الآلي في تطبيقات المكتبات	الحاسب
تطبيقات تكنولوجيا ؟ في المكتبات	المعلومات
استخدام الحاسب الآلي في تطبيقات؟	المكتبات
تطبيقات تكنولوجيا المعلومات في؟	المكتبات

ج. كشافات الكلمات الدالة المضافة للسياق

ويستخدم هذا النوع من الكشافات مع العناوين التي لا تتضمن مصطلحات كافية لوصف الوثيقة، حيث يقوم المكشف بإضافة كلمات تصف المحتوى الموضوعي للوثائق وعادة ما يستخدم في حالة العناوين المضللة أو العناوين القصيرة ويندر استخدام هذا النوع من الكشافات حالياً.

- **مميزات كشافات العناوين**

يتميز هذا النوع من الكشافات وكشافات التباديل بصفة عامة بما يلي:

1. سرعة وسهولة الإعداد

2. لا يحتاج إلى خبرة سواء موضوعية أو مهنية في إعداده.

3. انخفاض تكاليف إعداده.

4. ظهور المصطلحات الجديدة في التخصص الموضوعي بسرعة في هذا النوع من الكشافات، بحيث تصبح متاحة للبحث والاسترجاع، إلا أنه يتأثر بشكل واضح بعيوب اللغة الطبيعية كوسيلة لتكشيف وهي العيوب التي سبق ذكرها من قبل.

5.2.3.3 التكشيف الآلي
Automatic Indexing

يستخدم هذا الأسلوب في تكشيف أجزاء معينة من النص، لعل أبرزها تكشيف المستخلصات، حيث وجد أن المستخلص، خصوصاً مستخلصات المؤلفين تحتوي عدداً قليلاً من الكلمات، إلا أنها تحتوي على أكبر قدر من المعلومات الواردة في الوثيقة، كما أنها تصف بإيجاز محتوى الوثيقة.

ويتم إعداد هذا النوع من خلال تمييز كلمات المستخلص من خلال نظام التكشيف الآلي مع استبعاد الكلمات الواردة في قائمة الاستبعاد. ثم تكشيف كلمات المستخلص وفقاً للإجراءات التي تم عرضها عند الحديث عن نظم اللغة الطبيعية. وتتميز نظم التكشيف الآلي بمجموعة من الملامح الخاصة نذكر منها ما يلي:

1. بالطبع يمكن استخدام التكشيف الآلي في تكشيف النصوص الكاملة للوثائق وهو النمط السائد حالياً في معظم نظم استرجاع النصوص الكاملة وبعض النظم العاملة على شبكة الإنترنت.

2. تسمح نظم التكشيف الآلي أيضاً بعرض النتائج بأساليب عدة منها تقسيم النتائج المسترجعة إلى فئات فيما يعرف بـ Results Categorization، كما تسمح بتوجيه استفسارات ذات طبيعة خاصة مثل الاستفسارات التي

تتطلب إجابات على أسئلة Question Answering Query، كما تسمح أيضاً بالاسترجاع ما بين اللغات Cross Language Retrieval.

وقد أدى ظهور شبكة الإنترنت وخاصة الشبكة العنكبوتية إلى ظهور أنماط وطرق جديدة للتكشيف منها استخدام أساليب تحليل الروابط وتحليل نصوص الروابط في عمليات التكشيف الآلي وهو ما ستتعرض له بالتفصيل عند الحديث عن التكشيف والفرز على الويب.

المصادر

- حسام الدين، مصطفى (1996). محاضرات غير منشورة في استرجاع المعلومات.
- جلغوم، عبدالله (2012). مقدمة المعجم المفهرس الشامل لألفاظ القرآن الكريم بالرسم العثماني. ملتقى أهل التفسير، مسترجعة من الويب في 14 / 8 / 2018 https://vb.tafsir.net/tafsir 34016/# .W3JvVegzZPY
- لانكستر، ولفرد (1997) أساسيات استرجاع المعلومات/ ترجمة حشمت قاسم، الرياض: مكتبة الملك فهد الوطنية، 454 ص.

- Aitchison, J., Bawden, D., & Gilchrist, A. (2003). Thesaurus construction and use: a practical manual. Routledge.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.
- Luhn, Hans Peter. "Key word in context index for technical literature (kwic index)." American Documentation 11.4 (1960): 288-295.
- National information standards organization. (1993). Guidelines for the construction, format, and management of monolingual thesauri(ANSI/NISO Z39.19-1993). Bethesda, MD: NISO press.
- Rowley, J. (1992). Organizing knowledge: an introduction to information retrieval. Gower.
- Wellisch, H. H. (1995). Indexing from A to Z, 2nd. New York: HW Wilson

الفصل السادس

لغات تمثيل واسترجاع المعلومات في العصر الرقمي

◄ 6 مقدمة

تمت مناقشة الملامح والخصائص المميزة لكل من اللغة الطبيعية واللغة المضبوطة في الفصل السابق. ويستكمل هذا الفصل مناقشة قضية اللغة في تمثيل واسترجاع المعلومات في البيئة الرقمية مع التركيز على المراحل التي مرت بها لغات تمثيل واسترجاع المعلومات، والقضايا المتعلقة باللغة الطبيعية وأهميتها في البيئة الرقمية، ثم يستعرض الفصل مجموعة من لغات التمثيل الجديدة في البيئة الرقمية.

◄ 6.1 تطور لغات تمثيل واسترجاع المعلومات

بالنظر إلى تاريخ نظم تمثيل واسترجاع المعلومات تُعد اللغة المضبوطة أكثر حداثة في الاستخدام والتطبيق من اللغة الطبيعية، حيث كانت اللغة الطبيعية هي اللغة الأساسية في التواصل والتمثيل والوصف على مر العصور. وقد مرت عملية تطوير لغات التمثيل بأربع مراحل أساسية هي:

المرحلة الأولى: ترجع تلك المرحلة إلى العصور التي سبقت ظهور أي لغة اصطناعية مضبوطة وذلك حتى بداية القرن العشرين، حيث كانت اللغة الطبيعية هي اللغة الوحيدة المطبقة في كل نظم تمثيل واسترجاع المعلومات. وقد بدأ المستخدمون في تلك المرحلة إدراك القيود والمشكلات التي تنتج عن استخدام تلك اللغة مثل عدم الثبات في التعبير، الناتج عن مشكلات اللغة الطبيعية التي سبق عرضها، والتي تشمل المترادفات والمشترك اللفظي.

المرحلة الثانية: شهدت تلك المرحلة ظهور أول لغة مصطلحات مضبوطة والتي

تمثلت في تطوير خطط التصنيف كنموذج للربط المسبق. كما ظهرت أيضاً قوائم رؤوس الموضوعات والمكانز في النصف الأول من القرن العشرين. وقد بدأ في هذه المرحلة ظهور الجدل حول استخدام اللغة الطبيعية مقارنة باللغة المضبوطة في عمليات تمثيل واسترجاع المعلومات.

المرحلة الثالثة: شهدت عودة اللغة الطبيعية لتصدر المشهد مرة أخرى، كنتيجة لتطور نظم الاسترجاع التي تعتمد على الكلمات المفتاحية والنصوص الكاملة. واستمر تطبيق اللغات المضبوطة في تمثيل واسترجاع المعلومات في النظم الببليوغرافية مثل فهارس المكتبات في هذه المرحلة، ومع استمرار استخدام نظم اللغة الطبيعية لمعالجة النصوص الكاملة والمصطلحات المضبوطة لتمثيل واسترجاع النظم الببليوجرافية واحتدام الجدل حول أفضلية كل لغة ومزاياها وعيوبها ظهرت العديد من دراسات المقارنة بين اللغات لتحديد أفضل البدائل. وانتهت معظم هذه الدراسات إلى أن كل نظام له مزاياه وعيوبه.

المرحلة الرابعة: بدأت تلك المرحلة مع ظهور واجهات بحث اللغة الطبيعية في عمليات الاسترجاع، وقد استمرت اللغة المضبوطة مستخدمة في تلك المرحلة، ولكن في المشهد الخلفي فقط، حيث لم تعد تلك اللغات مرئية للمستفيد. وقد أطلقت عليها ميلستد (Milstead, 1995) المصطلحات المضبوطة غير المرئية في بيئة استرجاع المعلومات باللغة الطبيعية. وقد أسهمت التطورات المتلاحقة في نظم معالجة اللغة الطبيعية في تحقيق ذلك، ما أدى إلى ظهور نظم تعتمد بالكامل على اللغة الطبيعية مثل نظم (West Law and Lexis Nexis).

وعلى الرغم من عدم وجود حدود فاصلة قطعية بين المراحل الأربع التي مرت بها لغات تمثيل واسترجاع المعلومات؛ إلا أنه يمكن القول إن هذه اللغات قد تخطت المرحلتين الأولى والثانية، وما زالت تعمل في المرحلتين الثالثة والرابعة.

6.2 لماذا نحتاج إلى اللغة الطبيعية والمضبوطة معاً

يوجد نوعان أساسيان من لغات التكشيف هما (قاسم 2000):

- التكشيف بالتعيين: ويقصد به الجهد الفكري الذي يبذله المكشف في التحقق من عناصر المحتوى الموضوعي للوثيقة ثم اختيار المصطلحات أو المداخل الكشفية التي تعبر عن هذه العناصر، وذلك بالاعتماد على قوائم رؤوس الموضوعات أو خطط التصنيف أو المكانز.
- التكشيف بالاشتقاق: وفيه يتم اشتقاق أو اقتباس جميع المصطلحات من الوثيقة التي يتم تكشيفها وذلك بالاعتماد فقط على اللغة الطبيعية.

إن الاستمرار في الاعتماد على اللغتين كأساليب لتمثيل واسترجاع المعلومات، لابد أن يكون وراءه أسانيد دعت إلى ذلك، ولعل أبرز وأهم الأسانيد والأدلة هو وجود مزايا وعيوب لكل منهما، والتي أبرزتها دراسات المقارنة المستمرة حتى وقتنا هذا. ويمكن إيجاز تلك المزايا والعيوب في قدرة كل لغة من لغات التكشيف على معالجة إحدى القضايا التالية:

6.2.1 قضية المترادفات

الترادف هي المشكلة التي تنبع من إمكانية التعبير عن موضوع معين بعدة طرق مختلفة في وثائق مختلفة أو من جانب مكشفين مختلفين، ما يعني وجود أكثر من مصطلح واحد للدلالة على موضوع أو مفهوم معين. مثال لذلك: إذا أردنا التعبير عن مفهوم مثل التلفزيون نجد العديد من المصطلحات الدالة على هذا المفهوم مثل تلفزيون، تلفاز، تي في.. الخ أو أردنا استخدام مصطلح واحد مقنن للتعبير عن مفهوم التليفون المحمول يوجد العديد من المصطلحات المتداولة أيضاً مثل المحمول، الموبايل، النقال، الجوال، الخلوي وغيرها. ولا يمكن بأي حال من الأحوال استخدام كل هذه المصطلحات للتعبير عن مفهوم واحد عند استخدام اللغة المضبوطة، بالتالي لابد من الاختيار بينها. كما أنه لا يمكن للمستفيد أو

الباحث أن يتذكر كل هذه المصطلحات عند البحث، ما يظهر الحاجة إلى لغة مقيدة تضبط المصطلح المستخدم وتحيل إليه من الأشكال غير المستخدمة.

وتُعد قضية المترادفات إحدى أهم القضايا الجدلية التي تناولتها دراسات استرجاع المعلومات؛ حيث تشير معظم تلك الدراسات إلى أن القدرة على معالجة المترادفات أحد أهم عيوب اللغة الطبيعية. وعلى الجانب الآخر عند استخدام اللغة المضبوطة في عمليات تمثيل واسترجاع المعلومات، فإن قضية المترادفات تتم معالجتها من خلال اختيار مصطلح واحد للدلالة على كل المترادفات في عمليات التمثيل والاسترجاع، مع بناء نظام محكم للإحالات من المصطلحات غير المستخدمة إلى المصطلحات المستخدمة. ويُطلق على المصطلح المستخدم هنا للدلالة على المفهوم أو الكيان المصطلح المفضل Prefered Term والمصطلحات غير المستخدمة يطلق عليها الكلمات غير المفضلة Nonprefered Term.

6.2.2 قضية المشترك اللفظي

تظهر قضية المشترك اللفظي نتيجة لظاهرة يطلق عليها تعدد المعاني، والتي تُعد أيضاً من أبرز القضايا الجدلية في مجال المقارنة بين استخدام اللغة الطبيعية في مقابل اللغة المضبوطة. والمشترك اللفظي يدل على المصطلحات التي تحمل الشكل نفسه في الكتابة والهجاء وتختلف في المعنى. ويتم اكتشاف المعنى من خلال السياق الذي يرد فيه المصطلح.

وعادة ما يتم ترجمة مصطلح المشترك اللفظي إلى مصطلحين باللغة الإنجليزية هما Homonymy and Polysemy حيث يشير الأول إلى مجموعة من الكلمات لا علاقة بينها سوى اتفاقها في الصيغة والشكل (الجناس التام)، والثاني هو تعدد المعنى للكلمة وهو أقرب إلى المشترك.

ومن أمثلة المصطلحات التي تحمل مشتركاً لفظياً وتتنوع معانيها وفقاً للسياق الذي ترد فيه: جبن، جُبن؛ شعر، شِعر؛ عين (بيت) عين الإنسان، عين الماء؛ علم Science، علم Flag.

ويُعد سيبويه (ت 180 هجري) أول من أشار إلى قضية المشترك اللفظي، حيث ذكره في تقسيمات الكلام في كتابه قائلاً: «اعلم أن من كلامهم اختلاف اللفظين لاختلاف المعنيين، واختلاف اللفظين والمعنى واحد واتفاق اللفظين واختلاف المعنيين واتفاق اللفظين والمعنى مختلف». كما أفرد بن فارس (ت 395 هجري) للمشترك اللفظي باباً خاصاً وعرفه بقوله «معنى الاشتراك» أن يكون اللفظ محتملاً لمعنيين أو أكثر (محمد علي بيضون، 1997).

بالتالي، الاشتراك اللفظي مشكلة ناتجة عن غياب التحكم في اللغة، وتعني وجود كلمات متشابهة في الشكل ولكنها مختلفة في المعنى، أي الكلمات المتطابقة في الهجاء والمختلفة في الدلالة. ويصورها (لانكستر 1997) على النحوي التالي:-

مفهوم مثل عطارد Mercury نجد له العديد من المعاني مثل:

- شخصية أسطورية (إله التجارة والفصاحة عند الرومان)
- مصطلح (Mercury) يدل على كوكب سيار (عطارد)
- معدن الزئبق
- طراز سيارات

وتستطيع اللغات المضبوطة التميز بين المصطلحات المشتركة لفظياً من خلال استخدام تبصرات تحدد المعنى أو المجال بين قوسين مثال:

عطارد (أساطير)

عطارد (سيارات)

عطارد (معدن)

عطارد (كوكب)

ولا تقتصر قضية المشترك اللفظي على اللغة العربية ولكنها تظهر أيضاً في الإنجليزية، حيث يوجد الكثير من المصطلحات التي تشترك في البنية الحرفية، ولكنها تدل على أكثر معنى في اللغة الإنجليزية، ولا يفرق بينها سوى السياق الذي

وردت فيه مثل Record, subject, drug, spring, duty,.Bank……etc

وتؤدي ظاهرة المشترك اللفظي أو تعدد المعاني إلى غموض في الدلالة الاصطلاحية عند التمثيل والاسترجاع بسبب ضعف السياق أو التكشيف والبحث باستخدام كلمات مفردة. فعلى سبيل المثال إذا قام باحث باستخدام مصطلح مثل شعر في عملية البحث بصورة مستقلة، من الممكن للنظام أن يسترجع عدداً كبيراً من الوثائق التي ليس لها علاقة بالمعنى الدلالي الذي يقصده الباحث. ويرجع ذلك إلى أن المصطلحات عادة ما تكون غامضة في حد ذاتها ويزول عنها الغموض عندما يتم ربطها بغيرها من المصطلحات وعند وضعها في سياق محدد. وقد أشار كل من لانكستر وورنر (Lancaster and warner,1993) إلى مشكلة الغموض في الاسترجاع وهي عادة مشكلة نظرية أكثر منها مشكلة عملية، ذلك أنه نادراً ما تجد باحثاً يبحث عن كلمة مستقلة منفردة (عادة ما تكون غامضة) ولكنه عادة ما يربطها بكلمات أخرى تزيل الغموض عنها.

وتعتمد اللغة المضبوطة على أساليب متنوعة للتغلب على مشكلة المشترك اللفظي؛ حيث يتم تفسير المعنى المقصود للمشترك اللفظي باستخدام الهوامش التي ترد بين قوسين ()،() لتخصيص المعنى السياقي للمشترك اللفظي مثل:

عين (عضو الإبصار)

Duty [tax]

.[duty [responsibility]

◄ 6.2.3 قضية البحث الشامل

تنتج هذه المشكلة عن غياب التحكم في اللغة، ما يضطر المسؤول عن إجراء البحث إلى البحث بكل المصطلحات المتصلة دلالياً حتى يمكنه استرجاع كل أشكال ومرادفات المصطلح. بالتالي يسترجع كل النتائج الممكنة. وعادة ما تجمع اللغات المقيدة هذه المصطلحات المتصلة ببعضها بعضاً، إما هرمياً، كما هو الحال في خطط التصنيف، وإما دلالياً، كما هو الحال في المكانز وقوائم رؤوس الموضوعات.

◄ 6.2.4 قضية البنية

لكل لغة بنيتها الخاصة، ولكن كيف يمكن التعبير عن تلك البنية عند اختيار اللغة الطبيعية لتمثيل واسترجاع المعلومات؟ نفترض مثلاً أنه توجد وثيقة تم تمثيلها بثلاث مصطلحات باللغة الطبيعية هي: USA الولايات المتحدة الأمريكية، Automabiles السيارات، اليابان .Japan فهذه الوثيقة من الممكن أن تكون عن تصدير السيارات اليابانية لأمريكا أو عن تصدير السيارات الأمريكية لليابان. ويتبيّن أنه مع عدم وجود بنية واضحة لعلاقة المصطلحات توضح البناء اللغوي، يصبح من الصعب تحديد أي دولة هي التي تُصدر للأخرى عند استخدام تلك المصطلحات الثلاثة في تمثيل الوثيقة، من دون أي معلومات أخرى عن البناء اللغوي (بناء الجمل).

هذه المشكلة يمكن التغلب عليها بسهولة باستخدام رموز الأدوار في اللغة المضبوطة، وهي عبارة عن رمز أو رقم يحدد العلاقة البنائية syntax Relationship بين المصطلحات. ففي المثال السابق يمكن أن نستخدم رقم (1) للدلالة على المصدر ونضعه بعد المصطلح اليابان (1) بهذا الشكل (1) Japan للدلالة على أن اليابان هي المصدر. كما يمكن أن تخصص الرقم (2) للدلالة على الدور الثاني وهو المستورد وتخصصه لأمريكا (2) أو (2) USA. وتساعد هذه الرموز التي تسمح بها اللغة المضبوطة على معالجة قضية الخلط الذي يظهر نتيجة التداخل في البناء اللغوي، والتي لا يمكن معالجتها في اللغة الطبيعية.

◄ 6.2.5 قضية الدقة

تسعى كل نظم تمثيل واسترجاع المعلومات إلى استخدام لغة تستطيع التمثيل والبحث بدقة وفعالية. ومن الواضح أن هذا الهدف يمكن تحقيقه باستخدام اللغة الطبيعية في تمثيل واسترجاع المعلومات لسببين رئيسين هما:

الأول: أنه لا توجد أي معالجة إضافية مثل الشرح أو التعبير باستخدام الهوامش والإحالات عند استخدام اللغة الطبيعية في التمثيل والاسترجاع.

الثاني: أنه لا توجد حاجة إلى التفسير في اللغة الطبيعية، حيث إن المصطلحات

التي يتم البحث بها من جانب المستفيد هي نفسها مصطلحات اللغة المستخدمة في التمثيل والاسترجاع.

وعلى الجانب الآخر فإن اللغة المضبوطة هي لغة اصطناعية وهي أقل ثراءً من اللغة الطبيعية في تمثيل الوثائق واستفسارات المستفيدين. كما أن اللغة المضبوطة أقل تخصيصاً وتفتقر إلى التحديد الدقيق، ويرجع ذلك إلى إجراءات معالجة اللغة. ويبدو أن تفسير مصطلحات اللغة المضبوطة أمر لا مفرَّ منه؛ حيث إن المفهوم أو المعنى الدلالي لكل مصطلح يتم تحديده لخدمة نوعية معينة من المستفيدين، وقد يؤدي هذا التفسير إلى عدم الدقة في تمثيل واسترجاع الوثائق التي تعتمد على اللغة المضبوطة.

6.2.6 قضية التحديث

تعد قضية التحديث من أبرز مزايا اللغة الطبيعية، نظراً لأنها لغة ديناميكية تعتمد على المصطلحات التي ترد بالوثائق، من ثم فهي دائمة التحديث دون تدخل بشري في إجراء عملية التحديث. وفي المقابل تحتاج اللغة المضبوطة إلى التحديث الدائم والذي يعد أبرز عيوب اللغة المضبوطة، حيث إنها تتقادم بمجرد صدورها ويزداد معدل تقادمها يومياً. فالمصطلحات الجديدة تحتاج إلى أن يتم استخدامها في التمثيل والاسترجاع بمجرد ظهورها في الإنتاج الفكري، بينما تحتاج تلك المصطلحات الجديدة إلى إضافة وتحديد علاقات وإحالات وتدقيق حتى يتم إدراجها في اللغة المضبوطة، والتي تمر بعملية تحديث طويلة من حيث الوقت وصرامة الإجراءات. وينتج عن ذلك أن مصطلحات اللغة المضبوطة عادة ما تكون متقادمة، بينما يتم تحديث مصطلحات اللغة الطبيعية بصفة دائمة، ما يجعل الاستفسارات التي تحتوي على مصطلحات جديدة تواجه صعوبة في استرجاع الوثائق الصالحة عند استخدام اللغة المضبوطة، بينما يتم استرجاع الوثائق الحديثة والقديمة التي تشتمل على تلك المصطلحات بمجرد سك المصطلح واستخدامه في تمثيل الوثائق واسترجاعها باستخدام اللغة الطبيعية.

◄ 6.2.6 قضية الكُلفة

عـادة مـا تسـتغرق عمليـة بنـاء وصيانـة وتعليـم اسـتخدام اللغـة المضبوطـة وقتـاً طويـلاً فـي تمثيـل واسـترجاع المعلومـات، ويتـم ترجمـة ذلـك الوقـت المسـتغرق فـي هـذه الأنشـطة إلـى كُلفـة فـي نظـم تمثيـل واسـترجاع المعلومـات. وعلـى الجانـب الآخـر فـإن اللغـة الطبيعيـة هـي اللغـة التـي يسـتخدمها النـاس فـي التواصـل فيمـا بينهـم، مـن ثـم فهـي لا تتطلـب أي كلفـة إضافيـة؛ حيـث لا تحتـاج إلـى تدريـب أو صيانـة عنـد اسـتخدامها فـي تمثيـل واسـترجاع المعلومـات.

◄ 6.2.7 قضية التوافق

تظهـر الحاجـة إلـى تحقيـق التوافـق بيـن اللغتيـن المضبوطـة والطبيعيـة فـي بعـض الأحيـان فـي نظـم تمثيـل واسـترجاع المعلومـات، عندمـا تدعـو الحاجـة إلـى تغييـر اللغـة المسـتخدمة فـي النظـام أثنـاء تطويـره أو عندمـا يحتـاج المسـتفيد إلـى إجـراء البحـث فـي أكثـر مـن قاعـدة بيانـات فـي الوقـت نفسـه. لذلـك تظهـر قضيـة التوافـق فـي نظـم اللغـة المضبوطـة نظـراً لأن كل لغـة مـن اللغـات المضبوطـة لهـا ملامحهـا وخصائصهـا المميـزة لهـا. فعلـى سـبيل المثـال قـد يكـون مـن المسـتحيل اسـتخدام خطـة تصنيـف فـي إجـراء البحـث بالفهـارس المتاحـة علـى الخـط المباشـر بـدلاً مـن قائمـة رؤوس الموضوعـات (مكتبـة الكونجـرس). فـي حيـن أنـه عندمـا يتـم بنـاء نظـام اعتمـاداً علـى اللغـة الطبيعيـة فإنـه لا توجـد حاجـة إلـى التوافـق عنـد التغييـر، حيـث إن اللغـة الطبيعيـة مسـتقلة ومتوافقـة مـع نفسـها مـن حيـث البنيـة الاصطلاحيـة ومـن حيـث البنيـة الرمزيـة أيضـاً (لا توجـد رمـوز مسـتخدمة خـارج إطـار اللغـة بحروفهـا وكلماتهـا التـي تحمـل دلالات معينـة). وعـادة مـا يطلـق علـى هـذه القضيـة مصطلـح التشـغيل التبادلـي المسـتخدم فـي مجـال الحاسـبات الآليـة (Zeng & Chan,2004).

ويمكـن تلخيـص مزايـا وعيـوب كل لغـة فيمـا يلـي: نقـاط قـوة ومزايـا اللغـة المضبوطـة تتمثـل فـي معالجـة المترادفـات والمشـترك اللفظـي والبنـاء اللغـوي، والتـي تُعـد أيضـاً مـن أهـم عيـوب اللغـة الطبيعيـة. وبالمثـل فـإن نقـاط ضعـف اللغـة المضبوطـة تتمثـل فـي الدقـة والتحديـث والكُلفـة والتوافـق، والتـي تُعـد نقـاط قـوة وتميـز اللغـة الطبيعيـة.

وقد أشار رويلي (Rowley, 1992, 116) إلى ما يلي:

«يوجد اتفاق عام على ضرورة استخدام كل من اللغة الطبيعية والمضبوطة معاً، كما يوجد اتفاق عام على أهمية كل منهما في تمثيل واسترجاع المعلومات بأي نظام. وبعبارة أخرى أن كلاً منهما له أهميته في نظم تمثيل واسترجاع المعلومات. ولكن هل سيظل الأمر هكذا في المستقبل؟ هذا السؤال مازال مفتوحاً ولم تتم الإجابة عليه بسهولة في بيئة الويب الذكي والدلالي، إلا من خلال تطوير أدوات تجمع ما بين اللغتين».

6.3 لغات تمثيل واسترجاع المعلومات في العصر الرقمي

تم استخدام اللغتين المضبوطة والطبيعية بالتوازي في نظم تمثيل واسترجاع المعلومات في عالم مصادر المعلومات المطبوعة. ومازال التدخل البشري في التمثيل والاسترجاع قائماً في عالم مصادر المعلومات المتاحة على الخط المباشر، ما أعطى اللغة المضبوطة مكاناً ثابتاً في تلك البيئة. أما في العصر الرقمي فإن اللغة الطبيعية أصبحت النموذج الأساسي لتمثيل واسترجاع المعلومات، ونادراً ما تستخدم اللغة المضبوطة أو تستخدم على نطاق أضيق بكثير من استخدام اللغة الطبيعية. ويرجع ذلك إلى وجود العديد من الملامح المميزة للمعلومات في البيئة الرقمية، لعل أبرزها ما يلي:

- أن الغالبية العظمى من المعلومات الرقمية متاحة على الإنترنت في صورة نصوص كاملة، إلا أنها تفتقر إلى المراجعة والفحص، ما يعني الغياب الكامل لآليات ضبط الجودة.
- دورة حياة المعلومات في هذه البيئة قصيرة جداً، حيث تتغير المعلومات بسرعة كبيرة وبديناميكية مستمرة.
- تنمو المعلومات في ذلك الفضاء الرقمي بسرعة كبيرة وبمعدلات أُسية مضاعفة.

لذلك أصبح من الصعب تبرير استخدام اللغة المضبوطة المكلفة من حيث الوقت والمال في تلك البيئة التي تتسم بالديناميكية العالية والتغير السريع. من ثم اعتمدت معظم أنظمة استرجاع المعلومات الشهيرة المتاحة على الإنترنت (محركات بحث الويب) في تنفيذ مهام تمثيل واسترجاع المعلومات على اللغة الطبيعية، ولم تستخدم مطلقاً اللغة المضبوطة، بينما اعتمد عدد قليل من تلك النظم على قوائم الكلمات Word Lists، والتي تعد أقرب نموذج لاستخدام اللغة المضبوطة في تمثيل واسترجاع المعلومات في بيئة الإنترنت.

وعلى الرغم من ذلك، فإن اللغة الطبيعية لا يجب أن تكون اللغة الوحيدة في تمثيل واسترجاع المعلومات على الإنترنت، حيث إن ضعف الضبط الاصطلاحي قد يكون السبب الرئيس لعدم دقة النتائج التي يتم استرجاعها من نظم استرجاع الإنترنت. وبصفة عامة فإن مهمة الضبط الإصطلاحي قد تنتقل من على عاتق أخصائي المعلومات ويتحملها المستفيد النهائي عند استخدام اللغة الطبيعية في تمثيل واسترجاع المعلومات، حيث يحتاج المستفيد في العصر الرقمي إلى التفكير في المصطلحات المترادفة التي تتطلبها عملية البحث. فالتفاعل المتزايد والدائم بين المستفيد ونظم استرجاع المعلومات على الإنترنت سوف يُمكن المستفيد من أداء مهمة الضبط الاصطلاحي بفاعلية وكفاءة. ومع هذا التطور سوف يتحول دور أخصائي المعلومات من الوسيط في عملية البحث إلى المدرب على إجراءات البحث وكيفية الوصول إلى المعلومات، إلى جانب تقديم الدعم الفني للمستفيد في عملية البحث والاسترجاع. لذلك فالسؤال عن مستقبل الضبط الاصطلاحي في تمثيل واسترجاع المعلومات في البيئة الرقمية قد يكون من الصعب الإجابة عليه حتى الآن. مع ذلك فإنه توجد أربع طرق مختلفة لاستخدام الضبط الاصطلاحي في تمثيل واسترجاع المعلومات (Lancater & Warner,1994):

1. استخدام اللغة المضبوطة في كل من عمليات التمثيل والاسترجاع.
2. استخدام اللغة الطبيعية في كل من عمليات التمثيل والاسترجاع كوسيلة مساعدة على البحث والربط المسبق.

3. استخدام اللغة المضبوطة للتمثيل فقط، ويتم ضبط المصطلحات في عمليات الاسترجاع من خلال لغة مضبوطة مخفية أو ضمنية في النظام.

4. استخدام اللغة المضبوطة في عمليات الاسترجاع فقط، وقد تم تطبيق هذا النموذج في نظم يطلق عليها مكانز البحث فقط Search Only Theasaurus والتي يطلق عليها أيضاً الضبط الاصطلاحي اللاحق Post – Controlled Vocabulary.

وبالنظر إلى طبيعة وخصائص نظم تمثيل واسترجاع المعلومات في البيئة الرقمية، نجد أن البديل الثاني هو أكثر البدائل ملاءمة للتطبيق في تلك البيئة، حيث إن النموذجين الثالث والرابع يعملان على تخزين اللغة المضبوطة على الخط المباشر لدعم عملية البحث، والتي تبدو وكأنها بديل يمكن استخدامه لضبط المصطلحات عند الحاجة. ورغم ذلك فإن مجال تمثيل واسترجاع المعلومات قد شهد في السنوات الأخيرة ظهور مجموعة من اللغات الجديدة مثل التقسيم إلى الفئات، الفئات الاجتماعية، الأنطولوجيات. ومع أن لكل لغة من هذه اللغات ملامحها المميزة، فإن جميع هذه الأدوات تم تطويرها لأغراض التمثيل والاسترجاع في البيئة الرقمية.

6.3.1 علم التقسيم

تم مناقشة هذا المصطلح في الفصل الثاني باختصار، وتفصيلاً اشتق المصطلح Taxonomy من الإصل اليوناني taxis، والذي يعني الترتيب أو التصنيف ويستخدم المقطع nomos في الدلالة على القانون أو العلم. من ثم فإن المصطلح يشير إلى علم التقسيم إلى فئات أو علم االتقسيم. وقد استخدم المصطلح في بدايته في علم الأحياء للإشارة إلى تصنيف الكائنات الحية (الحيوانات والنباتات)، ثم اكتسب المصطلح دلالة أوسع من معناه الضيق في علم الأحياء، حيث يشير حالياً إلى تصنيف الأشياء، وامتد مفهومه إلى كل العلوم. وقد أشار جيلشرست (Gilchrist, 2003) إلى أن أول استخدام للمصطلح بمعناه الحديث ظهر سنة 1997 في مقالة عن ياهو Yahoo والذي يعد من أوائل أنظمة البحث في الإنترنت. وقد اشتهر بأنه أفضل دليل بحث استخدم نموذج التقسيم إلى فئات (أو علم االتقسيم).

وترجع جذور مصطلح علم التقسيم إلى خطط التصنيف والمكانز، فكما هو الحال في نظم التصنيف، تقوم أدوات (علم التقسيم إلى فئات)، بتعريف فئات محددة مسبقاً لإجراء عمليات التقسيم إلى فئات، وفقاً لقواعد علم التصنيف. وتعتمد نظم التقسيم إلى فئات على استخدام مستويات متنوعة من العرض - باستخدام النموذج الهجائي الرقمي alphanumenic؛ حيث لا تعتمد على نظام تصنيف محدد. ويتم التعبير عن العلاقات الترابطية بين الفئات باستخدام الترتيب الهجائي لكل مستوى، وذلك بمضاهاة أسلوب العرض والبناء الشائع في المكانز. وعلى خلاف خطط التصنيف والمكانز لا يستخدم علم التقسيم أي آلية أو نظام للإحالات، ما يضعف من وظيفته كنظام للضبط الاصطلاحي. وتعمل أدوات هذا النموذج على تيسير عملية التقسيم إلى فئات لدعم عمليات التصفح، والذي يُعد أحد أهم نظم الاسترجاع بعد البحث. ويُعد هذا النموذج فعالاً وجذاباً لمعلومات المؤسسات التي تسعى إلى بناء بوابات خاصة لتمثيل واسترجاع المعلومات، إلى جانب تطبيقاته في أدلة بحث الإنترنت (Gilchrist,2003).

ولعل أبرز أسباب استخدامه في بناء بوابات الشركات هو أنه نظام يساعد على استيعاب وتمييز المصطلحات التي تستخدمها الشركات والمؤسسات التجارية، إضافة إلى أنه أقل كُلفة من أي عملية بناء وصيانة لغة مضبوطة مثل المكانز. ويُعد دليل البحث ياهو (dir.yahoo.com) أبرز نموذج لبناء تلك الأدوات وأكثرها شمولاً على الإنترنت (Zhonghong, Chaudhry & Khoo 2006). وتجدر الإشارة إلى أن دليل البحث ياهو قد تم إغلاقه بعد عشرين عاماً من تشغيلة من 1994 حتى عام 2004، وتحول إلى بوابة بحث متكاملة تعتمد على محرك بحث وخدمات البوابات الإلكترونية والتي سيتم مناقشتها بالتفصيل لاحقاً. كما أن العديد من محركات البحث أغلقت أدلة بحثها بالكامل ومنها محرك البحث جوجل والذي تم تطويره في عام 2000 لمنافسة دليل ياهو، إلا أنه أُغلق في عام 2011. ولعل السبب الرئيس وراء إغلاق تلك الأدلة هو ارتفاع تكاليف تطويرها وصيانتها من جهة، وتطور إمكانات البحث بمحركات بحث الويب من جهة أخرى.

6.3.2 علم المصطلح الاجتماعي

تـم وصفـه فـي الفصـل الثانـي بأنـه العلـم الـذي يعتمـد علـى أسـاليب علـم التصنيـف التـي تتـم مـن خـلال تفاعـل الإنسـان مـع النظـام (Human System Interaction Vander, 2007).

ويقسـم بيتـر (Peters, 2009) علـم المصطلـح الاجتماعـي إلـى ثـلاث فئـات هـي كالتالـي:

- **الفوكسونومي الواسعة: Broad Folksonomy**

هـي أدوات تتيـح لمنشـئ المصـدر والمسـتفيدين الآخريـن إضافـة التعليقـات والكلمـات الدالـة علـى المصـدر سـواء كان (صـورة أو فيديـو.. إلـخ) أكثـر مـن مـرة.

- **الفوكسونومي الضيقة الممتدة: Extended Narrow Folksonomy**

وهـي الأدوات التـي تتيـح لمنشـئ مصـدر المعلومـات والمسـتفيدين الآخريـن التعليـق ولكـن لمـرة واحـدة فقـط. مثـال علـى ذلـك موقـع Flicker.

- **الفوكسونومي الضيقة: Narrow Folksonomy**

فـي هـذا النمـط يكـون مـن حـق منشـئ مصـدر المعلومـات فقـط إضافـة الكلمـات الدالـة والتعليقـات للمصـدر؛ ويكـون مـن حـق المسـتفيدين الآخريـن البحـث باسـتخدام هـذه الكلمـات فقـط. مثـال علـى ذلـك موقـع YouTube.

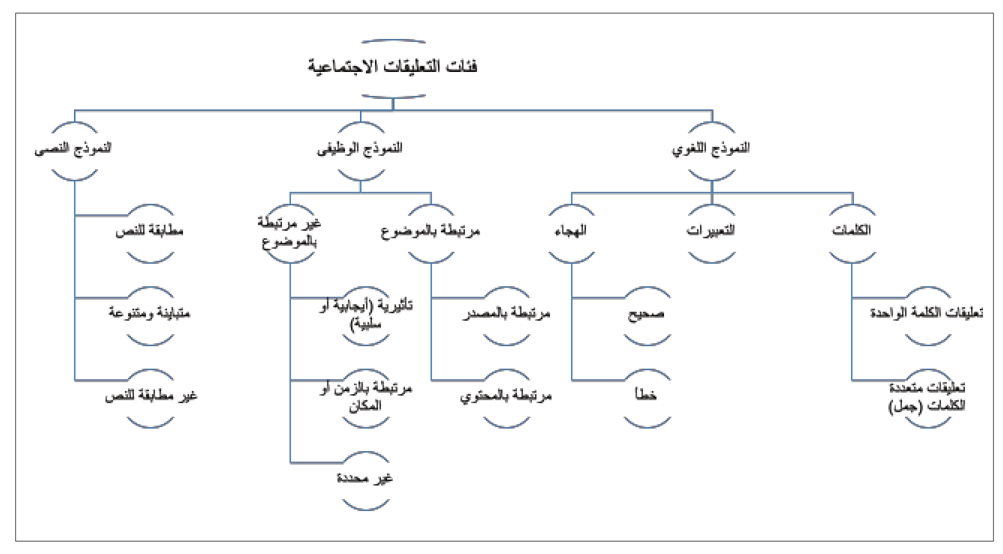
ويرتبـط علـم المصطلـح الاجتماعـي ارتباطـاً وثيقـاً بعمليـات التوسـيم الاجتماعـي Social Tagging والتـي تعـد أحـد مخرجاتـه الأساسـية، حيـث يتـم بنـاؤه بالاعتمـاد علـى التوسـيم الـذي يقـوم بـه المسـتفيدون أثنـاء عمليـات البحـث والتصفـح. وعـادة مـا تأخـذ المصطلحـات الاجتماعيـة شـكل سـحابة التوسـيم Tag Cloud والتـي تمثـل عرضـاً مرئيـاً لعمليـات التوسـيم التـي يقـوم بهـا المسـتفيدون. ويتـم اسـتخدام مصطلـح سـحابة التوسـيمات بديـلاً للمصطلحـات الاجتماعيـة أو مرادفـاً لهـا. وعلـى عكـس التصنيـف فـإن المصطلحـات الاجتماعيـة لا تظهـر لعـرض أي علاقـات هرميـة بيـن مكوناتـه (التوسـيمات). ويهتـم علـم المصطلحـات الاجتماعيـة بحفـظ العلاقـات الترابطيـة

Associative Relationship بيـن التوسـيمات ويقـوم بعرضهـا فـي ترتيـب هجائـي مـن دون إحـالات أو حواشـي مـن تلـك التـي يتـم تطبيقهـا فـي المصطلحـات المضبوطـة (مثـل المكانـز). مـن ثـم فـإن المصطلحـات الاجتماعيـة لا يمكـن معاملاتهـا بالطريقـة نفسـها الخاصـة بالمصطلحـات المضبوطـة، والتـي تمـت مناقشـتها سـواء مـن حيـث البنـاء أو التجميـع أو حتـى الوظيفـة. إضافـة إلـى ذلـك، فـإن كل نظـم المصطلحـات المضبوطـة، والتـي تتـراوح مـا بيـن خطـط التصنيـف إلـى علـم االتقسـيم (التقسـيم إلـى فئـات)، يتـم بناؤهـا بالاعتمـاد علـى أخصائـي المعلومـات، بينمـا يتـم بنـاء وتطويـر نظـم المصطلحـات الاجتماعيـة ـ والتـي تُعد نموذجاً جديداً للغـات تمثيـل واسـترجاع المعلومـات فـي البيئـة الرقميـة ـ بالاعتمـاد علـى المسـتفيد النهائـي ولصالحـه. وذلـك بغـرض الاسـتخدام فـي بيئـة الجيـل الثانـي للويـب 2.0 والتـي لا توجـد لهـا حـدود فاصلـة سـواء فـي الموضـوع أو الثقافـة أو حتـى الجغرافيـا (Munk & Mork, 2007).

وأثنـاء عمليـة التوسـيم يمكـن للمسـتفيدين أن يقومـوا باختيـار أي وسـم اصطلاحـي مـن المصطلحـات الاجتماعيـة المتاحـة، كمـا أنهـم يمكنهـم وضـع أو اختيـار أي وسـم اصطلاحـي جديـد مـن مصطلحاتهـم للدلالـة علـى الموضـوع الـذي يتـم وسـمه. ونظـراً لأن كل التوسـيمات فـي المصطلحـات الاجتماعيـة تكـون فـي صـورة روابـط فائقـة تُمكـن المسـتفيد مـن تصفـح المتـاح مـن التوسـيمات علـى المواقـع مـن خـلال روابـط التوسـيمات الفائقـة بجانـب إمكانيـة اسـتخدامها فـي البحـث. وقـد تـم مناقشـة مزايـا وعيـوب المصطلحـات الاجتماعيـة كلغـات لتمثيـل واسـترجاع المعلومـات بشـكل مكثـف فـي العديـد مـن الدراسـات والبحـوث ولعـل أبرزهـا: (e.g. Noruzi, 2006; Speller, 2007; Trant, 2006) سـواء مـن حيـث مقارنتهـا بعلـم التصنيـف والتقسـيم إلـى فئـات أو مـن حيـث علاقاتهـا بنظـم اللغـة المضبوطـة.

وبإيجـاز يمكـن القـول إن المصطلحـات الاجتماعيـة تحمـل كل مزايـا وعيـوب اللغـة الطبيعيـة مـع إضافـة ملمـح واحـد مـن ملامـح اللغـات المضبوطـة وهـو الترتيـب الهجائـي والعـرض المرئـي للتوسـيمات. مـن ثـم فـإن المصطلحـات الاجتماعيـة تعمـل وظيفيـاً كلغـة طبيعيـة أكثـر مـن كونهـا لغـة مضبوطـة فـي بيئـة تمثيـل واسـترجاع المعلومـات الرقميـة.

وقد لخص بيتر فئات التعليقات الاجتماعية (Peters, 2009, (p 203 في الشكل التالي:



شكل (6.1) فئات التعليقات الاجتماعية

6.3.3 الأنطولوجيات أو علم المصطلح الواحد

علم المصطلح الواحد أو الأنطولوجي استُخدم في مجال الفلسفة للدلالة على مفهوم دراسة الوجود. وقد سك المتخصصون في مجال الحاسب الآلي وخاصة الذكاء الاصطناعي مصطلح الأنطولوجي في عام 1980 للإشارة إلى تجميع وتمثيل المعرفة عندما يتم وضع إطار مفاهيمي لمجال معين أثناء تطوير النظم الخبيرة (Vickery , 1997).

ويتم تعريف مصطلح الأنطولوجيا في مجال هندسة المعرفة أو بشكل أوسع في علم الحاسبات والمعلومات على أنه عملية التخصيص الصريح والرسمي للأطر المفاهيمية المشتركة (Gruber, 1993). كما تم استخدامه للتعبير عن رؤية تيم بيرنر لي Tim Berenr Lee الخاصة بالويب الدلالي، حيث عدّه مكوناً أساسياً من مكونات رؤيته لبناء بيئة ويب تستطيع تمييز المعاني والدلالات من خلال الاعتماد على الأنطولوجيات (Berner – Lee , Henler & Lassila , 2001).

وقد وصف تيم لي الأنطولوجيات بأنها مجموعة من العبارات يتم كتابتها بلغة إطار وصف المصادر RDF والتي تحدد العلاقة بين المفاهيم وتضع قواعد منطقية لمسببات كل منها. ومن خلال متابعة الروابط التي تستخدمها الأنطولوجيات المخصصة تستطيع الحاسبات فهم المعنى الدلالي للبيانات التي تتضمنها صفحات الويب .(p.38)

ويوجد أشكال متنوعة للأنطولوجيات حصرها فيشولد (Vschold , 1996) في أربعة أشكال تتراوح ما بين غير الرسمية والرسمية الصارمة، وذلك من وجهة نظر هندسة المعرفة knowledge engineering وهي كالتالي:

النوع الأول: هو الأنطولوجيات غير الرسمية تماماً، والتي يتم التعبير عنها باستخدام لغة طبيعية فضفاضة.

النوع الثاني: الأنطولوجيات غير الرسمية ذات البناء structured informal ontologies وهي الأنطولوجيات التي توظف اللغة الطبيعية بطريقة محدودة وتحمل بنية واضحة بغرض تقليل الغموض وزيادة الوضوح في عرض المعرفة.

النوع الثالث: يطلق عليه الأنطولوجيات شبه الرسمية Semiformal Ontologies والتي يتم التعبير عنها باستخدام لغة اصطناعية محددة بشكل رسمي.

النوع الرابع: هو الأنطولوجيات الرسمية الصارمة Regorously formal outologics والتي تحدد المصطلحات بدقة باستخدام الدلالات الرسمية Formal sementic والنظريات المرتبطة بها.

وعلى الرغم من عدم وجود وصف واضح لنوع الأنطولوجيات المرتبطة ببيئة الويب الدلالي، إلا أن النموذج المحتمل للاستخدام في هذا المجال هو النوع الرابع المتمثل في الأنطولوجيات الرسمية كما أشار فيشولد (Vschold , 1996).

وتشتمل العلاقات بين المفاهيم التي تتضمنها الأنطولوجيات:

المترادفات synonymy.

المتضادات Antonymy.

المتشابهات hyponymy (التي تعبر عن العلاقات).

الجزئيات (الجزء) والتي تعبر عن علاقة الجزء (The Part of relation).

هـذه العلاقـات عـادة مـا يتـم اسـتخدامها فـي عـروض إطـار وصـف المصـادر RDF Graph والتـي تسـتخدم فـي بنـاء الويـب الدلالـي (2003), Grlchrist.

إضافـة إلـى ذلـك، فـإن الأنطولوجيات لابـد أن تحدد قواعـد منطقيـة للأسباب المتعلقة بالمفهـوم والعلاقـات المرتبطـة، والتـي تأخـذ شـكلاً ثابتـاً. علـى عكـس مـا يتـم فـي نظـم المصطلحـات المضبوطـة التقليديـة مثـل المكانـز، والتـي عـادة مـا تكـون العلاقـات فيهـا ثابتـة، فضـلاً عـن أنهـا يجـب أن يكـون بهـا آليات تعكـس التعبيـر المتواصـل عـن التحديثات التـي تتـم علـى المفاهيـم وإجـراء تلـك التحديثـات آليـاً. وتسـعى الأنطولوجيـات مـع غيرهـا مـن أدوات الويـب الدلالـي إلـى تحقيـق الفهـم للـدلالات والمعانـي التـي تحملهـا المعلومـات المتاحـة مـن مصـادر الويـب مـن خـلال أجهـزة الحاسـبات والبرمجيـات المستخدمة فـي تلـك البيئـة. عـلاوة علـى ذلـك فـإن وظيفـة الأنطولوجيات تختلـف بشـكل كبيـر عـن المصطلحـات المضبوطـة التقليديـة (المكانـز، خطـط التصنيـف.. الـخ)، حيـث إنهـا تسـتخدم لتحقيـق الفهـم الدلالـي لمصـادر الويـب باسـتخدام الحاسـبات وليـس تنظيـم عمليـات اسـتخدام المصطلحـات فـي نظـم تمثيـل واسـترجاع المعلومـات.

لقـد تطـورت الملفـات فـي العصـر الرقمـي بصـورة كبيـرة وتـم إجـراء العديـد مـن البحـوث والدراسـات فـي هـذا المجـال علـى الأدوات الجديـدة الملائمـة لتمثيـل الملفـات مثـل علـم التصنيـف (التقسـيم إلـى فئـات)، علـم المصطلـح الاجتماعـي، (التوسـيم الاجتماعـي) الأنطولوجيـات. كمـا تجـرى دراسـات حـول الانتقـال الاصطلاحـي Vocabulary switch والـذي يعـد طريقـة للتحـول الآلـي مـن لغـة تمثيـل واسـترجاع إلـى لغـة أخـرى بالمجـالات الموضوعيـة المختلفـة. ويُعـد هـذا التحـول مجـالاً خصبـاً لحـل مشـكلات، أو إنهـاء الجـدل الدائـر حـول اسـتخدام اللغـة الطبيعيـة أو اللغـة المضبوطـة، فبمجـرد تطبيقـه سـوف يصبـح لـدى المسـتفيد فرصـة الاختيـار بيـن اللغـة التـي يرغـب فـي

تطبيقها في عملية البحث، ولن يكون مضطراً إلى الالتزام أو محدوداً بنطاق لغوي محدد سواء كان مضبوطاً أو اصطناعياً، فضلاً عن إمكانية كسر الحواجز الموضوعة بين المجالات العلمية واستخدام كل المعلومات العلمية المتاحة بطريقة أكثر فعالية وكفاءة (Schatz, 1993). فالتحول الاصطلاحي يختلف تماماً عن استخدام الأنماط التقليدية للتحول المعروضة باستخدام لغة مضبوطة غير مرئية Invisible Controlled Vocabulary في أمرين أساسيين هما:

الأول: أن التحول الاصطلاحي يعتمد بكثافة على إجراء البحث باستخدام اللغة الطبيعية.

الثاني: التحول الاصطلاحي يتعامل مع لغات تمثيل واسترجاع المعلومات في العديد من المجالات (أي مجالات معرفية متنوعة)، بينما تتعامل المصطلحات المضبوطة المخفية أساساً مع الترجمة ما بين اللغتين الطبيعية والمضبوطة على الخط المباشر. فعلى سبيل المثال قام سشاتز Schatz بتجميع فضاء مفاهيمي Concept Space لعدد 10 ملايين مستخلص من مقالات الدوريات عبر أكثر من ألف مجال موضوعي تغطي مختلف قطاعات الهندسة والعلوم (Schatz, 1997)، وقد وجد أن هذه الفضائيات المفاهيمية أداة خصبة وفعالة لاقتراح التفاعل بين المصطلحات Interactive term Suggestion والتحول الاصطلاحي.

ويمكن القول باختصار إن عمليات التمثيل والاسترجاع الآلية مع استخدام الدلالات والفضائيات المفاهيمية تعد مستقبل معالجة اللغات في العصر الرقمي. وسوف يصبح هذا السيناريو حقيقة مع تحقيق رؤية تيم بيرنر لي ومساعديه للويب الدلالي.

المصادر

- بيضون، محمد علي (1997) الصاحبي في فقه اللغة العربية ومسائلها وسنن العرب في كلامها، لأحمد بن فارس بن زكريا القزويني الرازي، أبو الحسين (المتوفى: 395هـ)،، الطبعة الأولى، ص 59
- جلغوم، عبدالله (2012). مقدمة المحجم المفهرس الشامل لألفاظ القرآن الكريم بالرسم العثماني. ملتقى أهل التفسير، مسترجعة من الويب في 14 / 8 / 2018 https://vb.tafsir. net/tafsir34016 /W3JvVegzZPY
- قاسم، حشمت (2000). مدخل لدراسة التكشيف والاستخلاص. القاهرة: دار غريب. 317 ص
- لانكستر، ولفرد، وونر أرج. أساسيات استرجاع المعلومات: (نظم استرجاع – المعلومات) ترجمة حشمت قاسم. الرياض: مكتبة الملك الفهد الوطنية 1997 ـ 454 ص
- حسام الدين، مصطفى (1996). مجموعة محاضرات في استرجاع المعلومات، جامعة القاهرة.

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 34-43.
- Gilchrist, A. (2003). Thesauri, taxonomies and ontologies–an etymological note. Journal of documentation, 59(1), 7-18.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.
- Lancaster, F. W., & Warner, A. J. (1993). Information Retrieval Today. Revised, Retitled. Information Resources Press, 1110 North Glebe Rd., Suite 550, Arlington, VA 22201.
- Milstead, J. L. (1995). Invisible thesauri: the year 2000. Online and CD-Rom Review, 19(2), 93-94.
- Munk, T. B., & Mork, K. (2007). Folksonomy, the power law & the significance of the least effort. Knowledge organization, 34(1), 16-33.
- Noruzi, A. (2006). Folksonomies:(un) controlled vocabulary?. Knowledge organization, 33(4), 199-203.
- Rowley, J. (1992). Organizing knowledge: an introduction to information retrieval. Gower.
- Schatz, B. R. (1997). Information retrieval in digital libraries: Bringing search to the net. Science, 275(5298), 327-334.

- Speller, E. (2007). Collaborative tagging, folksonomies, distributed classification or ethnoclassification: a literature review. Library Student Journal, 2.
- Spiteri, L. F. (2007). The structure and form of folksonomy tags: The road to the public library catalog. Information technology and libraries, 26(3), 13-25.
- Trant, J., & with the participants in the steve. museum project. (2006). Exploring the potential for social tagging and folksonomy in art museums: Proof of concept. New Review of Hypermedia and Multimedia, 12(1), 83-105.
- Vander Wal, T. (2007). Folksonomy. http://www.vanderwal.net/essays/051130/folksonomy.pdf
- Vickery, B. C. (1997). Ontologies. Journal of information science, 23(4), 277-286.
- Zeng, Lei, M., & Mai Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. Journal of the American Society for information science and technology, 55(5), 377-395.
- Zhonghong, W., Chaudhry, A. S., & Khoo, C. (2006). Potential and prospects of taxonomies for content organization. Knowledge organization, 33(3), 160-169.
- Luhn, Hans Peter. “Key word in context index for technical literature (kwic index).” Journal of the Association for Information Science and Technology 11.4 (1960): 288-295.
- Peters, Isabella. Folksonomies. Indexing and retrieval in Web 2.0. Walter de Gruyter, 2009.

الفصل السابع

آليات الاسترجاع

وتمثيل الاستفسارات

مقدمة

يتناول هذا الفصل آليات البحث واسترجاع المعلومات والاعتبارات التي يجب مراعاتها عند إجراء عمليات البحث عن المعلومات، والتي تشمل تمثيل وصياغة الاستفسارات، إجراءات البحث وآلياته المختلفة سواء من حيث طريقة البحث أو حقول البحث، إضافة إلى آليات البحث المتقدم مثل البحث العشوائي، البحث الموزون، توسيع الاستفسارات، كما سيعرض الفصل أساليب اختيار آلية البحث الملائمة إلى جانب معايير تقييم نتائج البحث.

7 آليات البحث
Search Techniques

يتم تصميم آليات البحث المختلفة بغرض دعم المستفيد في الوصول إلى المعلومات التي يحتاج إليها بفاعلية وكفاءة. ومع التقدم الكبير الذي تشهده تكنولوجيا وبحوث ودراسات استرجاع المعلومات تتنوع وتتطور آليات البحث والاسترجاع. وعادة ما يتم تقسيم آليات البحث والاسترجاع إلى نوعين أساسيين هما: النموذج الأساسي والنموذج المتقدم.

7.1 آليات البحث الأساسية
Basic Search Techniques

يشتمل النموذج الأساسي على مجموعة آليات البحث البسيطة التي تشمل البحث البوليني، حساسية الحروف Case Sensitive، البتر، التقارب، البحث في

الحقول. وتدعم معظم نظم استرجاع المعلومات تلك الآليات بطرق مختلفة ومتنوعة وسيتم إلقاء الضوء على الملامح الوظيفية لكل نمط من تلك الأنماط عند إجراء البحث البسيط.

7.1.1 البحث البوليني
search Boolean

يُنسب المصطلح **بوليني Boolean** إلى عالم الرياضيات الإنجليزي **جورج بولي George boole** الذي طور طريقة التحليل الرياضي القائمة على المنطق **البوليني Boolean logic**. وقد استخدم بولي ثلاثة معاملات للتعبير عن المنطق البوليني وهي المعاملات التي تستخدمها كل نظم استرجاع المعلومات حتى اليوم وهي AND, OR, NOT وتشير AND إلى العلاقة (و) في اللغة العربية وتستخدم OR للتعبير عن العلاقة (أو)، أما NOT فتستخدم للتعبير عن علاقة الاستبعاد (ماعدا أو باستثناء).

ولتبسيط دلالات تلك المعاملات عادة ما يتم استخدام (AND) مع المفاهيم المتنوعة Different Concept لتشكيل علاقة بين مفهومين مختلفين أو أكثر، وتستخدم (ماعدا أو باستثناءNOT) لفصل أو استبعاد جزء صغير من المفهوم أثناء عملية البحث (Smith,1993)، بينما تستخدم أو (OR) لتضمين كافة الدلالات ضمن المفهوم الذي يتم البحث عنه، بحيث يتم استخدام المترادفات والأشكال المختلفة للمصطلح لتغطية كافة الصيغ التي ربما يرد بها المصطلح في الكشاف أو في النصوص عند إجراء البحث. وعند تطبيق تلك المعاملات في أي نظام استرجاع معلومات فإن النظام يفترض ما يلي:

- معامل الربط (و) AND يستخدم لتضييق نطاق البحث.
- معامل الحصر (أو) OR يستخدم لتوسيع نطاق البحث.
- معامل الاستثناء (ماعدا) NOT يطبق بغرض استبعاد النتائج غير المطلوبة والتي تدل على قطاع خارج نطاق اهتمام الباحث.

ويستخدم المعامل AND لدمج مصطلحين أو أكثر في عبارة البحث ويتطلب أن تكون كل المصطلحات المستخدمة في عبارة البحث موجودة في الوثيقة المسترجعة. فعلى سبيل المثال عبارة البحث: Filtering and Controversy تسوية ونزاع

يجب أن تسترجع وثائق بها المصطلحان معاً، بصرف النظر عن مكان ظهورهما في الوثيقة. وذلك بالاعتماد على آلية البحث وطرق إعداد الكشافات. ولن تسترجع هذه العبارة أي نتائج تتناول موضوعات ذات علاقة بتسوية النزاعات مثل المفاوضات السلمية، حظر الأسلحة نظراً لأنها لا تتطابق مع مصطلحات عبارة البحث. ويستخدم المعامل AND في البحث عن المفاهيم ذات العلاقة التي تشكل معاً مفهوماً أكثر تركيباً أو تعقيداً.

يستخدم معامل الحصر OR لتوسيع نطاق البحث من خلال تضمين مصطلحات لها أشكال متنوعة وذات علاقة بالمفهوم الرئيس الذي يتم البحث عنه. وعادة ما يستخدم المعامل OR في البحث عن المترادفات أو المصطلحات المرتبطة ببعضها بعضاً. ويتم استرجاع أي وثيقة تتضمن أي مصطلح من المصطلحات الواردة في عبارة البحث. فعلى سبيل المثال عبارة البحث السابقة تسوية النزاعات إذا تم استخدام المعامل OR في البحث عن المصطلحين كما يلي: Filtering OR Controversy تسوية أو نزاع، سوف تسترجع تلك العبارة أي وثائق بها مصطلح تسوية وأي وثائق بها مصطلح نزاع، كما أنها سوف تسترجع الوثائق التي ورد بها المصطلحان معاً. من ثم فإنه من الواضح أن المعامل OR يسترجع عدداً أكبر من النتائج التي يسترجعها المعامل AND لنفس العبارة ويساعد على توسيع نطاق البحث.

معامل الاستبعاد (ماعدا أو باستثناء) NOT هو معامل أكثر تعقيداً في عملية البحث إذا ما تمت مقارنته بالمعامل OR فعلى سبيل المثال: البحث عن العبارة التالية: Filtering NOT Controversy (التسوية NOT النزاع (سوف يسترجع كل الوثائق التي تتناول المصطلح تسوية وتستبعد الوثائق التي تتناول مفهوم النزاع، فعلى سبيل المثال سوف يتم استرجاع تنقية المياه، تنقية الهواء , Water Filtering Air Filtering ولكن سيتم استبعاد أي وثيقة تشتمل على المصطلح Controversy

من ثم فإن المعامل NOT يستخدم بغرض تحقيق عملية الاستبعاد للأجزاء والمفاهيم غير المرغوبة والتي يسعى المستفيد إلى استبعادها من نتائج البحث. ويتضح من ذلك أن المستفيد لابد أن يكون على دراية دقيقة باحتياجاته؛ لأن مصطلح تسوية باللغة العربية والإنجليزية يحمل دلالات متنوعة يحددها المفهوم الذي يبحث عنه المستفيد.

يطلق على عملية البحث باستخدام معامل واحد للربط عملية البحث البسيط Simple Search وفي حالة استخدام معاملين أو أكثر في عملية البحث يطلق عليها البحث المركب Compound Search وعادة ما يتم ترتيب أولويات البحث عند إجراء بحث بوليني متعدد المعاملات وفقاً للترتيب التالي:

- **أولاً معامل الاستبعاد NOT**
- **ثانياً معامل الربط AND**
- **ثالثاً معامل الحصر OR**

فعلى سبيل المثال عند إجراء البحث المركب عن العبارة التالية Filtering OR Censorship AND Controversy NOT Libraries (المصطلح Filtering يستخدم هنا بمعنى استبعاد) بالتالي يتناول الاستفسار السابق موضوع: الاستبعاد أو الرقابة والنزاع باستثناء المكتبات، سيتم إجراء عملية الاستبعاد من البحث أولاً، أي سيتم استبعاد أي وثيقة تشتمل على المكتبات من كل الوثائق التي تشتمل على المصطلح استبعاد. من ثم فإن النظام سيبحث أولاً عن الوثائق التي تشتمل على المصطلح استبعاد، ويستبعد منها كل الوثائق التي تشتمل على المصطلح مكتبات، ثم تجري علاقة الربط AND لاسترجاع كل الوثائق التي تشتمل على المصطلحين Censorship AND Controversy الرقابة والنزاع، حيث تسترجع كل الوثائق التي ورد بها المصطلحان، وأخيراً يتم الجمع بين المجموعة الأولى التي تضمنت الوثائق التي ورد فيها مصطلح استبعاد والتي استبعد منها، وكل الوثائق التي ورد بها مصطلح المكتبات، والمجموعة الثانية التي تم الربط فيها بين المصطلحين الرقابة والنزاع باستخدام المعامل (أو OR) من ثم يمكن الترتيب كالتالي:

المجموعة الأولى Filtering NOT Libraries

المجموعة الثانية Censorship AND Controversy

المجموعة الثالثة نتائج المجموعة الأولى OR المجموعة الثانية

وإذا لم تلب نتائج عملية البحث احتياجات المستفيد يمكنه وضع المصطلحات بين أقواس لتغيير الترتيب الطبيعي لعملية البحث أو تحديد الترتيب الذي يرغب أن تتم على أساسه العملية. ففي المثال السابق يمكن للمستفيد أن يقوم بوضع أقواس لتغيير الترتيب على النحو التالي مثلاً:

Fltering OR Censorship AND Controversy NOT Libraries (AND) ونتيجة لهذا التغيير في ترتيب أولويات الربط والاستبعاد والحصر ستجري عملية البحث وفقاً لترتيب الأقواس في العلاقات الرياضية التقليدية، حيث تبدأ عملية البحث بالمعامل OR يليه المعامل AND ثم المعامل NOT. مع العلم أن العلاقات الرياضية تتطلب فك الأقواس أولاً، حيث يتم فك القوس الأول (Filtering OR Censorship) للحصول على المجموعة الأولى ثم يتم فك القوس الأكبر. ثم يتم البحث في نتائج المجموعة الأولى بالربط مع AND Controversy المجموعة الثانية وأخيراً يتم استبعاد المكتبات من نتائج المجموعة الثالثة. من ثم تكون النتائج المسترجعة عن التسوية أو الرقابة المرتبطة بالنزاع باستثناء المكتبات. فكما هو واضح يمكن استخدام أكثر من قوس واحد لتحديد ترتيب معين في المعالجة بعبارات البحث المركب. لذلك عادة ما يطلق على البحث البوليني المركب مصطلح البحث المتداخل Nested Search.

وتعد آلية البحث البوليني أكثر وأهم آليات البحث التي تستخدمها كافة قواعد البيانات البليوجرافية على وجه الخصوص، سواء كانت فهارس مكتبات متاحة على الخط المباشر أو قواعد بيانات ببليوجرافية. ويتطلب إتقان عملية البحث البوليني التدريب الكافي على تراكيب المصطلحات وعلاقاتها ببعضها بعضاً والتعرف الدقيق إلى نظام تغطية كل قاعدة بيانات أو أداة البحث التي يتم استخدامها فى استرجاع المعلومات. وستتم مناقشة البحث البوليني ومقارنته بآليات البحث في محركات البحث في الفصل العاشر.

◄ 7.1.2 البحث الحساس (حساسية الحروف)

توجد العديد من اللغات التي يؤثر شكل كتابة الحروف في آلية البحث والنتائج المسترجعة، حيث تشتمل على الحروف كبيرة Upper Cases والحروف الصغيرة Lower Cases. ومن أمثلة تلك اللغات الإنجليزية والفرنسية والإسبانية. تسمح تلك الآلية للمستفيد بأن يحدد بدقة شكل كتابة الحروف بالمصطلحات التي يتضمنها الاستفسار وكيفية إرسالها لنظام البحث.

فعلى سبيل المثال المصطلح الإنجليزي Target باستخدام حرف T الكبير والمصطلح Target يمثلان نموذجاً بارزاً للكلمات التي تحمل معاني مختلفة مع الحروف الكبيرة والصغيرة. فالمصطلح Target يشير إلى مؤسسة بيع بالتجزئة وهو علامة تجارية شهيرة، بينما مصطلح Target يشير إلى الهدف أو المستهدف، بالتالي لا بد من أن يكون المستفيد على دراية أو وعي كاملين بالتمثيل الاصطلاحي وشكل كتابة الحروف الحساسة في المصطلحات التي تتطلب ذلك، حيث إن لها معاني مختلفة. من ثم يستطيع المستفيد في تلك الحالة أن يحدد ما إذا كان بحاجة إلى تحديد دقيق لشكل الكتابة أم يقتصر على الشكل التقليدي. فإذا كان المستفيد بحاجة إلى البحث عن مؤسسة البيع بالتجزئة التي تحمل العلامة التجارية Target فإنه في هذه الحالة لا بد أن يكتب المصطلح باستخدام حرف T الكبير. أما إذا كان المستفيد يبحث عن المصطلح بمعنى Target الهدف أو المستهدف فإنه يجب استخدام المصطلح في حالته بالحروف الصغيرة. وتجدر الإشارة إلى أن التطبيقات التي تستخدم هذا النموذج محدودة وقليلة جداً عند مقارنتها بالنموذج البوليني. وذلك على الرغم من أن البحث بالحروف الحساسة يساعد على إنجاز نوع معين من البحث والاسترجاع لا يمكن لأي آلية أخرى أن تحققه. مع العلم أن النموذج التقليدي لإجراء هذا النوع من البحث هو وضع بين أقواس الاقتباس « » من ثم إذا كان المستفيد بحاجة إلى Target العلامة التجارية فيمكنه وضع المصطلح بين قوسين عند إجراء البحث «Target» وسيفهم النظام أن المستفيد يبحث عن المصطلح بهذا الشكل، كما هو وسيستبعد كل المصطلحات التي تستخدم الشكل الصغير للحرف t في المصطلح Target.

وقد اعتمدت الكثير من نظم استرجاع المعلومات على آليات التطبيع في البحث Search Normalization والذي يؤدي إلى التوحيد وعدم التمييز بين الحروف الكبيرة والصغيرة، تركت مهمة التميز للمستفيد من خلال الاعتماد على سياق بحثي أو عبارة بحثية أكثر دلالة عن الموضوع. فيما استخدمت نظم أخرى آليات التقسيم إلى فئات، والتي تميز بين المعاني المختلفة للمصطلحات.

وتجدر الإشارة إلى أن مشكلة الحروف الحساسة تظهر بصورة أكثر وضوحاً في حالات معالجة المتشابهات في اللغة العربية، سواء حالات الجناس أو المشترك اللفظي، والتي تتطلب أن يكون النظام قادراً على معالجة تشكيل الحروف والتميز بين الأشكال المختلفة للكلمة من خلال التشكيل. وأبرز مثال لذلك عندما نبحث في محرك البحث جوجل عن كلمة «جُبن» يسترجع المحرك النتائج التالية:

طريقة عمل جُبن

فيديو يكشف خِسة وجُبن العناصر الإرهابية

بالطبع يتضح من السياق أن المفهوم الوارد في النتيجة الأولى يختلف عن المفهوم الوارد في النتيجة الثانية، على الرغم من الاشتراك اللفظي التام في شكل الكلمة بين النتيجتين.

ويتضح مما سبق أن مشكلة الحروف الحساسة يقع العبْء الأكبر فيها على المستفيد، وهي مجال خصب لبحوث الذكاء الاصطناعي ومعالجة اللغة الطبيعية.

7.1.3 البتر Truncation

يُعرّف البتر بأنه القطع أو الاجتزاء ويوجد العديد من المصطلحات المستخدمة للإشارة إليه مثل البدل Wildcard الجذع Stemming التجريد Stripping قناع المصطلح Term Mask أو خوارزمية التضاريس Conflation Algorithem. وتشير كل تلك المصطلحات إلى استرجاع الأشكال المختلفة للمصطلح، وذلك باستخدام جزء شائع أو عام بين كل تلك الأشكال المختلفة. وعادة ما تستخدم نظم استرجاع

المعلومات رمزاً مميزاً لعملية البتر مثل علامة الاستفهام ؟ أو النجمة * لتوجيه النظام إلى ضرورة استرجاع كافة الأشكال المختلفة للمصطلح. فعلى سبيل المثال عند البحث بالمصطلح network* فإن ذلك يُعد توجيهاً للنظام باسترجاع كل الكلمات الأخرى للمصطلح مثل networking , networks, networkable ... الخ. يوجد ثلاثة أنماط أساسية للبتر هي:

- **النوع الأول بتر اللواحق Suffix** والذي عادة ما يطلق عليه البتر الأيمن Right Truncation والذي يعد الممارسة الأكثر شيوعاً في عمليات البتر، مع مراعاة أشكال الكتابة المختلفة بين العربية والإنجليزية.
- **النوع الثاني يُطلق على بتر السوابق Prefix** والذي يقوم ببتر الأجزاء الأولى من المصطلحات ومثال على ذلك *graduate من الممكن أن تشير إلى المصطلح Postgraduate, Undergraduate.Semigraduate ويطلق على هذا النوع البتر الأيسر Left Truncation وهو نادر الاستخدام ولا توجد أنظمة تقريباً تدعمه في العصر الحالي وعادة ما يترك لفهم المستفيد.
- **النوع الثالث هو البتر الأوسط Infix Truncation** ويشير إلى بتر أجزاء من وسط الكلمة، وأحياناً يطلق عليه البتر الداخلي. وتجدر الإشارة إلى أن البتر الأوسط أحياناً يستخدم علامة الاستفهام (؟) في الإشارة إلى عدم تأكد المستفيد من الحرف المحذوف أو رغبة المستفيد في استرجاع الأشكال المختلفة لهجاء الكلمات. فعلى سبيل المثال عند استخدام المصطلح clo?r عند إجراء البحث فإن النظام سوف يسترجع المصطلحات Color, Colour، كما أن البحث باستخدام Organi?ation سوف تسترجع Organization AND Organisation. وعادة ما يطلق على عملية البتر الأوسط مصطلح البحث بالحروف البديلة Wildcard.

ويمكن القول إن البتر يساعد المستفيد على استرجاع الأشكال المختلفة للمصطلح باستخدام الشكل الشائع وتحديد مواضع الاختلافات. ويجب على المستفيد أن يحدد الجزء الشائع في المصطلح وأماكن الأجزاء التي يوجد بها اختلافات. وعلى

الجانب الآخر يجب عدم الإسراف في عمليات البتر لأجزاء كبيرة من المصطلح؛ حيث إن بتر مصطلح مثل catalog إلى * cat يؤدي إلى استرجاع كم كبير من الوثائق غير الدقيقة عن القطط مثلاً، وعلى الجانب الآخر فإن بتر عدد أقل من اللازم من الحروف قد يُفقد المستفيد فرصة استرجاع وثائق مهمة. فعلى سبيل المثال استخدام الشكل catalogu كنموذج لبتر المصطلحات الدالة على مفهوم الفهارس سوف يضيع على المستفيد فرصة استرجاع وثائق تستخدم المصطلح الأمريكي catalog في مقابل استرجاع وثائق تستخدم الشكل البريطاني catalogue، ولتحقيق بعض التحكم في عملية البتر تسمح بعض النظم بتحديد عدد الحروف التي يتم بترها.

7.1.4 البحث بالتقارب
Proximity Search

يعمل المعامل البوليني AND على تحديد المصطلحات التي يجب أن تتضمنها الوثيقة المسترجعة؛ إلا أنه لا يحدد المسافة بين تلك المصطلحات ومدى تقاربها من بعضها بعضاً. فعلى سبيل المثال عبارة البحث البولينية Filtering AND Controversty (النزاع AND التسوية) قد تسترجع وثائق تتضمن مصطلحات بجوار بعضها بعضاً، أو متباعدة مئات الكلمات عن بعضها بعضاً، أو في أي مكان بالوثيقة مثل أن يظهر أحد المصطلحات في عنوان الوثيقة والآخر في نهاية الوثيقة. وقد يؤدي ذلك إلى أنه لا توجد علاقة على الإطلاق بين تلك المصطلحات المسترجعة، ما يؤدي إلى استرجاع وثائق لا تتناول الموضوع الذي يبحث عنه المستفيد، ولحل تلك المشكلة تم ابتكار أسلوب بحث يعتمد على تحديد مدى التقارب بين المصطلحات ومدى الارتباط بينها في إطار سياق معين عادة ما يطلق عليه البحث بالتقارب أو البحث بالتجاور Adjacency Search.

ويسمح البحث بالتقارب للمستفيد أن يحدد بدقة مدى التقارب أو المسافة بين المصطلحات البحثية وعلاقاتها الموضوعية Relative Position باستخدام المعامل مع with والمعامل بالقرب near. وتختلف تلك المعاملات من نظام إلى نظام آخر.

ويشير المعامل with إلى أن المصطلحين المستخدمين في البحث لا بد أن يظهرا بجوار بعضهما، كما وردا وبنفس الترتيب المستخدم في العبارة البحثية؛ فعلى سبيل المثال، العبارة البحثية Iformation with Technology تشير إلى أن الوثائق المسترجعة لهذه العبارة لا بد أن تتضمن العبارة Information Technology كما هي وليس أي شيء آخر مشابهاً مثل Information and Technology أو Technology and Information، إضافة إلى ذلك، فإنه يمكن تحديد عدد الكلمات التي تفصل بين المصطلحات عند استخدام المعامل with حيث يتم إضافة عداد (N) لتحديد عدد الكلمات التي تفصل بين المصطلحين المستخدمين في البحث N with ويتم استبدال N بعدد الكلمات (1,2,3...) الفاصلة بين المصطلحين وتحديد ترتيب تلك المصطلحات.

فعلى سبيل المثال العبارة البحثية information 2 with technology تسترجع وثائق عن

Information technology

Information and technology

Information and network technology

Information retrieval technology

من ثم فإن هذه العبارة البحثية سوف تسترجع الوثائق التي ترد فيها المصطلحات المحددة بالعبارة البحثية على مسافة لا تتجاوز مصطلحين فقط.

كما يستخدم المعامل بالقرب near بنفس الطريقة التي تشير إلى أن المصطلحين الذين تم ربطهما ببعضهما بعضاً لابد أن يكونا متجاورين adjacent، ولكن على عكس المعامل **with** فإن المصطلحين المستخدمين مع المعامل **near** من الممكن أن يظهرا في أي ترتيب ما داما متجاورين في النص. على سبيل المثال العبارة البحثية information near technology تسترجع وثائق عن information technology أو technology information.

كما يستخدم المعامل بالقرب N near لتحديد عدد الكلمات التي تفصل بين المصطلحين المستخدمين في العبارة البحثية؛ حيث يتم تحديد عدد الكلمات

(1,2,3...) n بصرف النظر عن ترتيبها في الوثائق والعبارة البحثية، حيث يمكن أن يأتيا في أي ترتيب ظهرا فيه في الوثيقة، فعند البحث بالعبارة البحثية information near technology 2 يمكن للنظام أن يسترجع أياً من الوثائق التي تشتمل على المصطلحات التالية:

information and technology

information and networked technology

technology and information

technology and business information

ويُعد البحث بالجمل searching pharse النموذج الأكثر استخداماً في نظم استرجاع المعلومات الحالية للدلالة على البحث التجاوري، وعادة ما يستخدم مع النظم التي تتعامل مع الكلمات وتكشيف الكلمات words index. وبتحديد أكثر دقة فإن المعامل with يمكن أن يقوم بإجراء بحث بالجمل المتطابقة exact pharses search من حيث المصطلحات والترتيب عند البحث باستخدام المعامل near. كما يقوم بإجراء البحث عن الجملة البحثية بصرف النظر عن مواقع الكلمات أو ترتيبها، ولكنه يلتزم بمدى تقاربها كما وردت في العبارة البحثية. وتقوم بعض الأنظمة بتوسيع نطاق التجاور في عمليات البحث ليشمل التجاور في الحقول البحثية والتجاور في الفقرات بدلاً من تحديد عدد محدد من الكلمات.

فعلى سبيل المثال نظام ديالوج Dialog لاسترجاع المعلومات عن الخط المباشر يسمح للمستفيد بتحديد البحث التجاوري سواء باستخدام with or near في حقول بحثية محددة. وتجدر الإشارة إلى أن معظم نظم استرجاع المعلومات الحالية تعتمد بصورة أكبر على البحث بالجمل من خلال استخدام التعبير عن الجمل البحثية بين الأقواس المزدوجة « » وهو نمط مستخدم في قواعد البيانات ومحركات البحث المتاحة على الويب على السواء. وقد تخلت معظم تلك النظم عن تعقيدات البحث التجاوري باستخدام معاملات with and near واستبدالها بالأقواس المزدوجة في الدلالة على الجمل البحثية (pharse searching).

7.1.5 البحث في الحقول
Field Searching

تعد التسجيلات الببليوجرافية التي يتم إعدادها لتمثيل أوعية المعلومات من أهم أساليب التعبير عن شكل ومحتوى الوثائق. وتتكون أي تسجيلة ببليوجرافية من مجموعة من الحقول التي تمثل المؤشرات الأساسية لأوعية المعلومات. وتشمل الحقول الببليوجرافية بيانات عن المؤلفين والعناوين وبيانات النشر والموضوعات.. الخ. وعادة ما ينظر إلى الحقول على أنها الوسيلة الأساسية للدلالة على معلومات الوثيقة مثل المؤلف والعنوان، بيانات النشر ونوع الوثيقة.. إلخ. وعادة ما يتم تمثيل الوثائق من خلال تلك الحقول البحثية وهي المحددات الأساسية أو بدائل الوثائق في أي نظام استرجاع معلومات، بالتالي فإنه عندما يتم تمثيل الوثائق باستخدام حقول تمثل تسجيلات أو بدائل للوثائق يمكن استخدام نفس الحقول في البحث عن الوثيقة. ويساعد البحث في الحقول على تحديد عملية البحث في حقل معين أو مجموعة من الحقول. ويعمل البحث الحقلي على تحقيق وظيفتين أساسيتين هما:ـ

الوظيفة الأولى: تحديد الحقل الذي يرغب المستفيد أن تكون المعلومات التي يبحث عنها قد وردت فيه، فعلى سبيل المثال إذا كان المستفيد يبحث عن أعمال شخص معين مثل Hans Peter Luhn's المرتبطة بمجال استرجاع المعلومات information Retrieval من الممكن البحث باستخدام المصطلح استرجاع المعلومات في الموضوع، إلا أن ذلك سوف يسترجع عدداً كبيراً من الوثائق عن استرجاع المعلومات التي ألفها Hans Peter Luhn's وغيره في نفس الموضوع. أما إذا حددنا البحث باستخدام حقل المؤلف، فسوف يتم استرجاع كل وثائق المؤلف التي تناولت موضوع استرجاع المعلومات. مع العلم أنه قد تم الربط بين الحقليين البحثيين باستخدام المعامل البوليني AND.

الوظيفة الثانية: استخدام البحث الحقلي يساعد على تضييق نطاق البحث بفاعلية، نفترض أن باحثاً قام بإجراء بحث عن موضوع علم المعلومات Information Science فإن هذا النوع من العمليات البحثية سوف يسترجع عدة آلاف من الوثائق التي تتناول

الموضوع، وعدد قليل جداً من الباحثين سيكون لديهم القدرة والوقت على مراجعة كل تلك الوثائق، من ثم فإنه يمكن تضييق نطاق البحث بفاعلية في الموضوع نفسه في حقول مثل سنوات النشر، اللغة، نوع الوثيقة.

ويُعد البحث الموضوعي باستخدام الموضوعات subject أو المفاهيم concept أو المجالات topics والذي يطلق عليه البحث عن مضمون المعلومات aboutness of information متبوعاً بالبحث عن موضع المعلومات ofines of information الذي يتم تحديده من خلال الحقول البحثية هو الطريقة المثلى لإجراء البحث عن نتائج محددة. وتجدر الإشارة إلى أن معظم محركات البحث المتاحة على الإنترنت لا تتيح إمكانية البحث باستخدام الحقول، نظراً لأن المعلومات لا يتم تمثيلها باستخدام بدائل حقلية للتعبير عن محتوى الوثيقة، كما هو الحال في نظم استرجاع المعلومات التقليدية، لذلك فإن البحث الحقلي غير قابل للتطبيق في محركات بحث الإنترنت.

7.2 آليات البحث المتقدم
Advanced Retrieval Techniques

يتم تطبيق كل آليات البحث البسيط، في معظم، إن لم يكن كل، نظم استرجاع المعلومات، وفي المقابل يتم تطبيق آليات البحث المتقدم في عدد محدود واختياري من أدوات البحث والاسترجاع أو تستخدم في الاختبارات المعملية للمقارنة بين كفاءة النظم. وتوجد نماذج متنوعة للبحث المتقدم سيتم تناولها بالتفصيل في الجز التالي:

7.2.1 البحث الغامض
Fuzzy Searching

يطلق عليه أحياناً البحث المجرد وهو نمط من أنماط البحث يشبه البحث بالبتر truncation مع بعض الاختلافات الأساسية، فبينما يسمح البحث بالبتر باسترجاع الأشكال المختلفة للمصطلح من خلال تحديد الجزء المتشابه في عملية البحث ويضع علامة البتر عند الجزء المختلف أو المشكوك في صحته؛ فإن البحث

الغامض يُستخدم في الوصول إلى المصطلحات التي يوجد بها أخطاء هجائية سواء عند كتابة الاستفسار أو إدخال البيانات في النظام، فعلى سبيل المثال المصطلح computer من الممكن أن تحدث أخطاء هجائية عدة عند كتابته فيكتب compyter or compture compiter or cometer فيحتاج النظام إلى آلية لتصحيح تلك الأخطاء عند البحث عن تلك المعلومات، كما تظهر تلك المشكلة عند إجراء رقمنة لوثائق مطبوعة وتحويلها إلى نصوص باستخدام نظم التعرف الضوئي إلى الحروف Optical Character Recognition (OCR) إلى جانب النصوص المضغوطة compressed text التي تظهر بعض الأخطاء عند فك ضغطها Uncompress في بعض الحروف. وقد تم تطوير آلية البحث الغامض لتحديد وتصحيح أخطاء الهجاء التي تنتج عن أخطاء إدخال البيانات في التمثيل أو صياغة الاستفسارات أو الاختلافات في نظم التعرف الضوئي على الحروف أو النصوص المضغوطة (Grossman & Frieder,(1998 ويعد نموذج تكرار المصطلحات n-gram أحد أهم الآليات المتخصصة في تطبيق البحث الغامض. وهو عبارة عن وضع نماذج لتفكيك الكلمات بطول محدد يطلق عليه n gram متبوعاً بسلسلة من الحروف (n (2, 3, 4,... n في الكلمة أو أن يتم فك أو تحليل المصطلح إلى أجزاء حسب عدد n من الأجزاء. فإذا أخذنا المصطلح Fuzzy Searching كنموذج من الممكن أن تكون لدينا أساليب تحليل الثنائية والثلاثية التالية (Kowalski,(1997:

Bi-grams (n=2): fu uz zz zy

Se ea er rc ch hi in ng

Tri-(n=3): fuz uzz zzy

Sea aer arc rch chi hin ing

توجد أساليب تحليل الأجزاء (n-grams) الرباعية quart grams والخماسية penta grams وطرق أخرى تستخدم في الإجراءات التحليلية للاستفسارات وجودة إدخال البيانات والتحليل الصرفي لنظم التعرف الضوئي على الحروف والنصوص المضغوطة. هذا النمط التحليلي n-grams ليس من الضروري أن تكون له أي علاقة

بالمعنى الدلالي للمصطلح، على الرغم من ذلك فهو يستخدم بكثافة في نظم التدقيق الإملائي والتحقق من الأخطاء.

وتستخدم خوارزميات المضاهاة لتحديد ما إذا كان هناك تطابق بين طريقة التمثيل والاستفسار الذي يدخله المستفيد إلى النظام، فإذا كانت كل الأجزاء n-grams الخاصة بمصطلحات التمثيل مطابقة تماماً لمصطلحات الاستفسار لا يقوم النظام بأي عملية تصحيح، أما في حالة عدم تطابق جزء أو جزئين one-gram or two gram يقوم النظام بإظهار خطأ في الإدخال (Grossman & Frieder,1998).

وقد أصبح تطبيق البحث الغامض في معالجة الأخطاء أو اقتراح التصويبات الممكنة في الكثير من أنظمة البحث، ومن الأمثلة الشائعة أيضاً لتطبيق آليات البحث الغامض استخدامه في المقارنة بالقواميس، حيث يتم مقارنة كل كلمة بالاستفسار الذي يرد إلى نظام استرجاع المعلومات بأحد القواميس. وفي حالة تحديد أي خطأ بعملية الإدخال يتم تصحيح الخطأ من خلال المطابقة بالمصطلح القاموسي وتصحيحه. ويمكن القول في المجمل إن البحث الغامض يساعد الأنظمة على التغلب على مشكلات أخطاء إدخال البيانات سواء في عملية التمثيل أو الاستفسارات. من ثم فالوثائق التي تتضمن أخطاء نتيجة الأخطاء الهجائية أو عدم دقة نظم التعرف الضوئي على الحروف أو أخطاء فك الضغط وغيرها من الحالات المشابهة لم يكن من الممكن استرجاعها دون وجود آلية البحث المجرد.

7.2.2 البحث بوزن المصطلحات
:Term weighted searching

يعرف وزن المصطلحات بأنه عملية إعطاء قيمة أو وزن نسبي للمصطلح المستخدم في تمثيل الوثيقة و/ أو استفسار المستفيد. ففي بعض الأحيان يحتاج المستفيد إلى تسليط ضوء أكبر على بعض أجزاء الجمل البحثية أكثر من غيرها. فعلى سبيل المثال في العبارة البحثية Filtering AND Controversy النزاع والتسوية، قد يكون المستفيد أكثر اهتماماً بجانب النزاع منه بجانب التسوية، بالتالي فهو بحاجة إلى إعطاء وزن نسبي

للمصطلح نزاع أكبر من الوزن النسبي للمصطلح تسوية، ولهذا الغرض يتم تصميم نظم البحث بالوزن النسبي، بحيث يمكن تخصيص درجات أو قيم للمصطلحات يطلق عليها البحث بالوزن النسبي لاستفسارات المستفيدين، وذلك بغرض تحديد الأجزاء الأكثر أهمية التي تحتاج إلى تسليط الضوء عليها بصورة أكبر من الأجزاء الأقل أهمية.

ويتم تحديد الأوزان بصور مختلفة، منها وضع رمز مثل النجمة * بجوار المصطلح كما هو الحال في قاعدة بيانات ERIC للدلالة على أنه مصطلح أساسي أو باستخدام دلالات رقمية Numerals سواء كانت عشرية أو صحيحة. كما تستخدم بعض النظم نظام درجات من 1ـ5 لإعطاء نقاط تدل على الأهمية حيث تشير (5) إلى أعلى درجة و (1) إلى أقل درجة. وبالطبع فإن عملية إجراء البحث بالوزن النسبي تتطلب أن تكون عملية التمثيل نفسها قد وضعت أوزاناً للمصطلحات في مرحلة التمثيل. فعلى سبيل المثال عند إجراء البحث باستخدام العبارة البحثية (6) Filtering (3) AND Controversy النزاع (6) و التسوية (3) فإن المستفيد يتوقع أن النظام سوف يسترجع وثائق تشتمل على هذين المصطلحين بنفس الوزن النسبي، بحيث يكون وزن الوثائق المسترجعة للمصطلح نزاع تعادل 6 في حين يكون وزن المصطلح تسوية في الوثائق المتسرجعة يعادل 3.

ومن الممكن استخدام درجة قطع أو حد معين Threshold لتخصيص الوزن الذي يلبي احتياجات المستفيد. نفترض أنه تم تعيين الحد كالتالي (Controversy 3) AND Filtering (6) النزاع (6) والتسوية (3). فإن الحد هنا هو 9 درجات، من ثم فإن أي نتائج بحد أقل من (9) حتى لو كانت الوثيقة تتناول نفس الموضوعين بأوزان 3 للتسوية و6 للنزاع، فإنها سوف تعد وثيقة غير صالحة للاستفسار ولا تلبي الأوزان التي تم تحديدها في الاستفسار.

من الواضح أن عملية تحديد قيم أو درجات نسبية للمصطلحات هي المعيار الأساسي لآليات البحث بالوزن. توجد العديد من خوارزميات الوزن Weighting Algorithms المستخدمة في تحديد أوزان المصطلحات منها:

موضع المصطلح Term Location

تقارب المصطلح Term Proximity

تردد المصطلح (TF) Term Frequency

عكس تردد المصطلح (ITF) Inverse Documents Frequency

الأحكام الفردية Individual Judgements

وعلى الرغم من وجود كل تلك الخوارزميات التي يمكن أن تستخدم في وزن المصطلحات، إلا أن الأحكام الفردية للمستفيدين أو الطريقة الحتمية Determmenistic Method أو التحديدية يمكن تطبيقها بصورة عملية من جانب المستفيد، حيث يمكن للمستفيد في الوقت نفسه تحديد الأوزان الخاصة بالمصطلحات في العبارة البحثية، دون أن يكون على دراية بأوزانها في الوثائق. وفي المقابل فإن كل الأساليب الأخرى لتخصيص الأوزان تعتمد على وزن المصطلحات المشتقة من الوثائق التي يتم تكشيفها، لذلك فإن آليات الوزن التي تعتمد على موضع وتقارب وتردد المصطلح يمكن تطبيقها فقط مع نظم التكشيف بالوزن النسبي Weighted Indexing .

تعتمد نظم الوزن بالأحكام الفردية على أحكام ذاتية غير موضوعية من جانب المستفيد، إلا أن تطبيقها يعتمد على مزيج من العوامل التي تشمل الحاجة إلى المعلومات، وطبيعة نظم استرجاع المعلومات، وشكل النتائج المتوقعة من حيث الوزن. بعبارة أخرى، فإن المستفيد عندما يحدد وزن المصطلحات في الاستفسار يجب أن يراعي هذه العوامل عند إجراء البحث، لذلك فإن تخصيص الوزن في وقت بناء الاستفسار لا يعد بهذه الطريقة إجراءً اعتباطياً Arbitrary بصورة كاملة.

وكما سبقت الإشارة، توجد العديد من معايير تخصيص الأوزان التي تستخدم مع نظم التكشيف بالوزن النسبي للمصطلحات أكثر من نظم البحث بالوزن النسبي. ومن ضمن تلك المعايير خوارزميات موضع المصطلح والتي تشير إلى مكان ظهور المصطلح في الوثيقة، ووفقاً لتلك الطريقة فإن المصطلحات التي تظهر في مواضع معينة من الوثيقة يتم تحديدها مقدماً وتخصيص أوزانها وتكون أكثر أهمية من المصطلحات التي تظهر في أجزاء أخرى من الوثيقة ومن أبرز المواضع التي تركز عليها هذه النوعية من أنماط التكشيف (العناوين) رؤوس الأجزاء، والعناوين الجانبية.. إلخ.

وتشير خوارزمية تقارب المصطلحات إلى المسافة بين المصطلحات الكشفية في الوثيقة. وبصفة عامة كلما قلّت المسافة بين المصطلحين وتقاربا في الوثيقة، ارتفع الوزن النسبي لتكشيف تلك المصطلحات. بمعنى آخر أن المصطلحات المتقاربة تحصل على وزن نسبي أكبر من المصطلحات المتباعدة في الوثيقة، فعلى سبيل المثال يحصل مصطلح استرجاع المعلومات على وزن نسبي أكبر في الوثيقة عندما يراد استرجاع المعلومات أكثر من مصطلحات أخرى مثل استرجاع المعلومات والبحث عنها. ويحصل مصطلح نظم استرجاع المعلومات على وزن نسبي أعلى من نظم خزن واسترجاع المعلومات أو في التمثيل والبحث والاسترجاع المعلوماتي وهكذا.

وتستخدم خوارزمية تردد المصطلحات للتعبير عن عدد مرات ورود المصطلح في الوثيقة، فكما أوضحنا من قبل أن الكلمات التي ترددت كثيراً في الوثيقة ليس شرطاً أن تكون مصطلحات كشفية مهمة، نظراً لأنها قد تكون كلمات وظيفية Function Words أو كلمات تعبيرية Expression Word وليس لها أي دلالة اصطلاحية واسترجاعية بالوثيقة. وفي المقابل فإن المصطلحات التي يكثر ترددها في الوثيقة، والتي تعبر عن مصطلحات كشفية مهمة بالوثيقة لابد أن يتم إعطاؤها وزناً نسبياً مرتفعاً في الدلالة على مضمون الوثيقة.

وتجدر الإشارة إلى أن تلك المصطلحات تتردد بكثرة في وثائق معينة، ونادراً ما تتردد في بقية الوثائق بقاعدة البيانات، ما يساعد على التمييز بين الكلمات الوظيفية والتعبيرية وكلمات الوقف Stop Words والمصطلحات الكشفية كثيرة التردد في الوثائق الفردية المهمة التي تحصل على أوزان نسبية مرتفعة في تكشيف وتمثيل تلك الوثائق (Salton,1989).

وعند حساب تردد المصطلح في الوثيقة تتم مراعاة عدد الوثائق التي يرد بها المصطلح في تخصيص وزن المصطلح، ويعرف هذا المقياس بـ(مقابل وعكس تردد المصطلح (Inverse Document Frequency - idf). ففي منتصف الستينات من القرن الماضي، توصل العالم الأمريكي كليفردون C.W. Cleverdon إلى وسيلة لتحديد الوزن النسبي للمصطلح في الوثيقة بهدف تكشيف الوثائق بصورة أفضل. وكنتيجة لأعمال

كليفردون حاول من بعده العديد من الإحصائيين والرياضيين التوصل إلى خوارزمية لتحديد قيمة المصطلح ضمن مجموعة من الوثائق. وقد سعت التحليلات في البداية إلى التركيز على مفاهيم واختيارات لمجموعة من المصطلحات، وقد تطور الأمر بعد ذلك لاستخدام كل المصطلحات الواردة في الوثيقة لتحديد الوزن النسبي للمصطلح ضمن الوثيقة، ومن هنا جاء الاهتمام بخوارزمية (مقابل تردد الوثائق). ويتم قياس مقابل تردد الوثائق بحسابات لوغارتيمية Logarithmic Calculation وهو عبارة عن معدل النصوص والوثائق التي توجد ضمن المجموعة الكاملة للوثائق وعدد الوثائق التي تحتوي على المصطلح المحدد. من ثم فهي عبارة عن معدل لوغارتيمي لعدد الوثائق التي تشتمل على مصطلح ما إلى إجمالي عدد الوثائق بالنظام (spnrckjones,2000).

ويعني ذلك أنه كلما انخفض عدد الوثائق التي ورد بها المصطلح، ارتفع وزنه النسبي في التمثيل لهذه الوثيقة، وكلما ارتفع عدد الوثائق التي ورد بها المصطلح انخفض وزنه النسبي في تمثيل الوثيقة.

وعادة ما تستخدم خوارزمية تردد المصطلحات TF مع خوارزمية مقابل تردد المصطلحات idf ويطلق على هذه الخوارزمية تردد المصطلحات في مقابل تردد الوثائق Tf.idf. وفي أحيان أخرى يتم مراعاة طول الوثيقة DL (Document length) عند تطبيق خوارزمية تردد المصطلحات في مقابل تردد الوثائق، كمؤشر إضافي لتحديد وزن المصطلحات في الوثيقة. فعند تثبيت معدل تردد المصطلح وعدد الوثائق التي ورد بها المصطلح، فإنه كلما كانت الوثيقة أكثر طولاً من الوثائق الأخرى، كان المصطلح الذي ورد بها أقل أهمية من الوثائق الأقل طولاً. فمثلاً إذا ورد مصطلح 5 مرات في وثيقة طولها 1000 كلمة فهو أقل أهمية في هذه الوثيقة من مصلطح ورد 5 مرات في وثيقة طولها 100 كلمة.

وقد تم تطبيق خوارزمية حجم الوثيقة في تردد المصطلح في مقابل تردد الوثائق tf.idf في العديد من تجارب مؤتمر استرجاع النصوص (Text Retrienal conference-TREC) للمقارنة بين العديد من الأنظمة (spnrckjones,2000).

كما توجد العديد من آليات وزن المصطلحات الأخرى التي تم تطبيقها من جانب

مطوري النظم مثل الأساليب الاحتمالية Probability Appraach وأساليب الاستدلال Inferences Approach (والتي سيتم مناقشتهما لاحقاً)، إلا أنها تستخدم من خلال المزج بينها وبين طرق أخرى مثل موضع المصطلح، والتي يتم تطبيقها مع خوارزمية تقارب المصطلحات في خوارزميات وزن المصطلحات. وتجدر الإشارة إلى أن محركات بحث الإنترنت تستخدم آليات وزن المصطلحات من خلال وضع رموز وعلامات بجوار المصطلحات البحثية مثل (- ,+ ,*. « » الخ).

7.3 توسيع الاستفسارات

Query Expansion

توسيع الاستفسارات إحدى آليات الاسترجاع التي تتيح للمستفيد تحسين النتائج المسترجعة من خلال مراجعة الاستفسارات بناء على النتائج المسترجعة التي تعطى المستفيد انطباعاً عن مدى دقة صياغة العبارة البحثية. وتعد عملية توسيع الاستفسارات عملية تكرارية وتفاعلية حيث يقوم فيها المستفيد بتعديل العبارة البحثية من خلال مراجعته للنتائج المسترجعة في أكثر من دورة بحثية لنفس الاستفسار.

ففي إطار عملية توسيع الاستفسارات يتم فحص النتائج المسترجعة لاستخلاص المعلومات الدالة التي يمكن من خلالها إعادة صياغة الاستفسار، وعادة ما تتكرر تلك العملية من الناحية النظرية حتى يحصل المستفيد على نتائج مرضية، وينصح المستفيد في المراحل الأولى من البحث بقراءة كل العناوين والمستخلصات المرتبطة ببحثه حتى يستوعب كل المصطلحات الدالة على الموضوع وعلاقتها ببعضها بعضاً؛ حيث إن التفاعل المستمر بين المستفيد ونظام استرجاع المعلومات يساعد على تحسين النتائج من خلال تحسين مستوى إدراك المستفيد لمحتوى النظام.

وقد أشار كل من ريسنك وفاوخان (Resnick & Vaughan,2006) إلى وجود طريقتين للتعامل مع الاستفسارات في هذا السياق، الأولى هي توسيع الاستفسارات، أما الثانية فهي تضييق الاستفسار Query Expanding and Narrowing.

إذا كانت عملية توسيع الاستفسارات تتضمن إضافة المترادفات والمصطلحات

المرتبطة بعبارة البحث بغرض زيادة عدد النتائج الصالحة المسترجعة؛ فإن تضييق نطاق البحث يهدف إلى استخدام مصطلحات أكثر تحديداً أو استبعاد المصطلحات التي تحمل معاني متشابهة غير ذات علاقة بموضوع البحث. من ثم فإن التوسيع الغرض منه إضافة نتائج صالحة إلى قائمة النتائج المسترجعة، بينما التضييق الغرض منه استبعاد النتائج غير الصالحة من قائمة النتائج المسترجعة.

التوسيع عادة ما يضيف أو يوسع نطاق العلاقات الاصطلاحية المرتبطة، سواء في نفس المستوى الشجري لمصطلحات العبارة البحثية أو في المستويات الأعلى. أما التضييق فإنه عادة ما يستبعد مصطلحات أو يحدد المصطلحات بصورة أكثر دقة ويعمل على إزالة الغموض Disambiguty الاصطلاحي بغرض التأكد من استرجاع النتائج الصالحة فقط واستبعاد النتائج غير الصالحة.

ويتم تقسيم عملية توسيع الاستفسارات إلى ثلاث فئات بناء على مصدر اختيار المصطلحات المرتبطة بعملية توسيع الاستفسار (Gauch, Wang & Erachakonda,1999) وهي:

- التخصيص الاصطلاحي Term Specificaty وهو عبارة عن إجراء عملية توسيع بالاعتماد على مجموعة فرعية من الوثائق المسترجعة باستخدام استفسار أولي ثم مراجعة المصطلحات الواردة في الوثائق المسترجعة، بناء على تلك المجموعة الفرعية، ويطلق على تلك العملية التوسيع بتخصيص الاستفسار Query Specific Expansion. وإذا تمت عملية التوسيع بناء على مجموعة المصطلحات التي يتم تحديدها أو الحصول عليها من خلال تحليل محتوى قاعدة بيانات نصوص كاملة معينة، من ثم فإنها عملية تخصيص بناء على ذخيرة نصية Text Corpus Specific.

- التخصيص اللغوي Language Specificity من خلال البحث في الأدوات المضبوطة مثل المكانز وقوائم رؤوس الموضوعات العامة وغير المرتبطة بمجموعة محددة من الوثائق. ويمكن أن تتم عملية توسيع الاستفسارات بطريقة يدوية أو آلية. ويقوم المستفيد في الطريقة اليدوية بتحديد المصطلحات الجديدة وإجراء عملية تعديل الاستفسار بنفسه. أما التوسيع الآلي، والذي

يطلق عليه أيضاً رد فعل الصلاحية Relevance Feedback والذي يعتمد على افتراض أن مجموعة النتائج التي ترد على قمة الترتيب Top Ranked في نتائج البحث هي المجموعة الأكثر صلاحية، من ثم استخدامها في عملية مراجعة وتوسيع الاستفسار ولا يتدخل المستفيد سواء بطريقة مباشرة أو غير مباشرة في عملية تعديل الاستفسار (Grossman & Frieder ,1998, Salton,1990) .

وتجدر الإشارة إلى أن مصطلح توسيع الاستفسار ليس المصطلح الملائم لوصف تلك العملية، والمصطلح الأكثر دلالة هو تعديل الاستفسار Query Modifications. ومن الآليات الإضافية لتعديل الاستفسارات استخدام قوائم المقترحات، والتي يتم إدراجها في صورة قائمة منسدلة أثناء إجراء البحث، تقترح مجموعة من المصطلحات عندما يقوم المستفيد بإدخال الاستفسار في صندوق البحث.

وقد يرى البعض أن هذه الآلية قد تؤدي إلى تشتيت المستفيد User Distraction، إلا أن البعض الآخر يرى أنها تدعم عملية التوسيع في الوقت الحقيقي Real Time Expansion بمعنى أن عملية التعديل تتم بصورة تفاعلية مع استفسارات المستفيدين (White & Marchionini, 2006).

- ترتيب النتائج Results Ranking تعد عملية ترتيب النتائج وسيلة أساسية لتعديل الاستفسار من خلال استخدام أسلوب الصلاحية الراجعة في عملية التوسيع الآلي للاستفسار، كما هو الحال في آليات الوزن Weighting Techniques التي تعتمد على خوارزميات الوزن والترتيب مثل موضع المصطلح، تقارب المصطلحات، تردد المصطلحات.. الخ.

وتعتمد كل نظم استرجاع المعلومات على خوارزمية خاصة بالترتيب، عادة ما تكون غير منشورة أو متاحة للجمهور العام. ولعل أبرز الأساليب المستخدمة في الترتيب في بيئة الويب استخدام أسلوب شهرة الروابط Link Publarity ومنها الروابط الراجعة Back Link الذي يعتمد عليها محرك البحث جوجل منذ عام 1998 (Vidman,1998). وتعتمد تلك الطريقة في الحكم على صلاحية أي صفحة أو موقع ويب إلى جانب معايير أخرى بناء على عدد الروابط التي تشير إليها باستخدام الروابط الفائقة.

ومن الأساليب الأخرى المستخدمة في توسيع الاستفسارات استخدام نموذج الاستفسار بالمثال Query by Example، حيث يشير مثال هنا إلى النتائج التي يتم استرجاعها، من ثم يتم استخدامها كنموذج في الحصول على نتائج أخرى. ففي نظم البحث عن الأصوات والصور والوسائط المتعددة من الممكن أن يستخدم النموذج من المستفيد مباشرة مثل استخدام رسم باليد كنموذج Hand Drawn Sketch يقوم المستفيد بإدخاله إلى النظام، كما يمكن أن يقوم المستفيد بإدخال نغمة معينة للبحث عن الأصوات. وتعتمد العديد من نظم استرجاع المعلومات التي تعمل في بيئة الإنترنت على أساليب التوسيع من خلال علاقات التشابه والصلاحية الراجعة باستخدام الربط الفائق الذي يمكن للمستفيد النقر عليه مثل «أكثر من هذا» More Like This.

تعد عملية تعديل الاستفسار إحدى أهم آليات تحسين النتائج المسترجعة والتي تعتمد على مراجعة الاستفسار من خلال اقتراح مصطلحات في صناديق البحث أو بالاعتماد على نتائج مسترجعة بالفعل في الحصول على نتائج مثيلة لها. ويلعب هذا النموذج وخاصة الصلاحية الراجعة دوراً كبيراً في تحسين أداء أدوات البحث على الإنترنت، من ثم فإن له تطبيقات عدة في العصر الرقمي.

7.4 بحث قواعد البيانات المتعددة
Multiple Databases search

يستخدم مصطلح البحث في قواعد البيانات المتعددة أو البحث العام أو البحث المجمع في الإشارة إلى عمليات البحث في أكثر من قاعدة بيانات أو أداة بحث بالتزامن في الوقت نفسه. ويشير مصطلح قاعدة البيانات هنا إلى أي نظام استرجاع معلومات سواء كان محرك بحث أو فهرساً أو قاعدة بيانات.. الخ. ويتميز هذا النمط من أنماط البحث بثلاث مميزات أساسية هي:

1. أن البحث في نظام استرجاع معلومات واحد قد لا يسترجع كل النتائج التي يحتاج إليها المستفيد، نظراً لأن لكل نظام تغطيته الموضوعية ونقاط تركيزه

وملامحه الخاصة التي تختلف عن نظام آخر، وفي هذه الحالة لابد من توسيع نطاق البحث ليشمل كل المصادر المحتملة.

2. البحث المتعدد قد يساعد المستفيد على عملية اختيار المصدر الملائم للبحث، إذا كان المستفيد غير متأكد أو مدرك للنظام أو النظم الملائمة لاستفساره. فالمستفيد المبتدئ يمكنه أن يعتمد على البحث المتعدد للتعرف إلى المصادر المتاحة ثم الانتقال إلى مرحلة التحديد والفلترة من خلال التصفح..

3. النتائج التي يحصل عليها المستفيد من البحث المتعدد تساعده على التعرف إلى النظم الملائمة لإجراء بحث فيها في المستقبل، بمعنى أن البحث المتعدد يعمل هنا كنظام توصية Suggesting Systems.

عند إجراء البحث في قواعد البيانات المتعددة يجب على المستفيد أن يراعي الاختلافات في تراكيب الاستفسارات Query Syntax واللغة وقدرات البحث الخاصة بكل نظام من أنظمة استرجاع المعلومات المستخدمة في البحث المتعدد، حيث إن الملامح الأساسية والشائعة في أحد النظم قد لا تكون متاحة في نظم أخرى. كما أن الملامح والإمكانيات البحثية الشائعة في أكثر من نظام قد يتم التعبير عنها وتفسيرها بطرق مختلفة من نظام لآخر. فعلى سبيل المثال تستخدم قواعد بيانات المعامل البوليني AND، بينما تستخدم محركات البحث معامل الجمع (+) في الدلالة على عمليات الربط بين المفاهيم المتنوعة بغرض تحديد نطاق البحث. كما توظف العديد من قواعد البيانات المعامل AND على أنه الإعداد الافتراضي Default Setting لعمليات البحث عند الربط بين أي كلمتين. بينما يوظف عدد محدود جداً من قواعد البيانات الأخرى المعامل OR كإعداد افتراضي.

وتجدر الإشارة إلى أن اللغات المستخدمة في التكشيف بالنظم المتعددة في الغالب ما تكون غير متشابهة، فتوجد احتمالات لاستخدام اللغات الطبيعية وأخرى لاستخدام اللغات المضبوطة في قطاعات موضوعية مختلفة. ومن الصعوبات الأخرى التي تواجهها نظم البحث المتعدد هو كيفية معالجة أشكال البيانات المختلفة مثل: الشكل أسكي ASCII

لتمثيل البيانات والفهارس المقروءة آلياً MARC والتي يتم تخزينها في قواعد البيانات. مع العلم أنه يتم استخدام بروتوكول Z39.50 لخدمات استرجاع المعلومات وهو البروتوكول المخصص لتطبيقات المكتبات إلى جانب معايير أخرى كمعايير التشغيل التبادلي ومنها على سبيل المثال معيار RDF Resource Description Framework لمعالجة كل أشكال البيانات لأغراض الاسترجاع. لذلك فإن النظم التي تتوافق مع معيار Z39.50 يُمكن إجراء البحث المتعدد فيها بسهولة بصرف النظر عن الاختلافات في أشكال البيانات أو مدى تقاربها الجغرافي (Michael & Hinnebusch,1995).

ويعد معيار Z39.50 المعيار الأساسي المعتمد من جانب المؤسسة الوطنية لمعايير المعلومات National Information Standards Institute لتطبيقات فهارس المكتبات المتاحة على الخط المباشر OPAC وفهارس الويب WebPAC وغيرها من نظم استرجاع المعلومات من قواعد بيانات ببليوجرافية وقواعد بيانات نصوص كاملة. ويعتمد معيار Z39.50 على استخدام واجهة موحدة بصرف النظر عن الواجهة التي يستخدمها كل نظام على حدة.

ومع نمو متطلبات العمل في بيئة الويب ظهرت معايير جديدة للبحث والاسترجاع في هذه البيئة، منها خدمة البحث والاسترجاع من الويب (Search \ Retrieve WebSearvice -SRW) والبحث والاسترجاع من خلال معين المصادر الموحد (Search Retrieve Via URL -SRU). وقد تم تصميم هذين البروتوكولين لتيسير إجراءات البحث سواء إرسال الاستفسارات أو تلقي النتائج في بيئة الويب. فعندما يقوم المستفيد بإرسال استفسار عبر نظام بحث متعدد فإن تراكيب التعبير عن الاستفسار قد تختلف من نظام لآخر، كذلك شكل نتائج الاستجابة، حيث إن الاستجابة لا تقتصر فقط على نتائج البحث ولكن أيضاً على شكل المعلومات Formatting Information. وأحياناً يتم الدمج بين البروتوكولين SRW/SRU معاً في بروتوكول واحد يتم الإشارة إليه بالمختصر (SRW\U) والذي يقوم بمعالجة مشكلات التراكيب المتنوعة والاستجابات المختلفة في نظم البحث المتعدد.

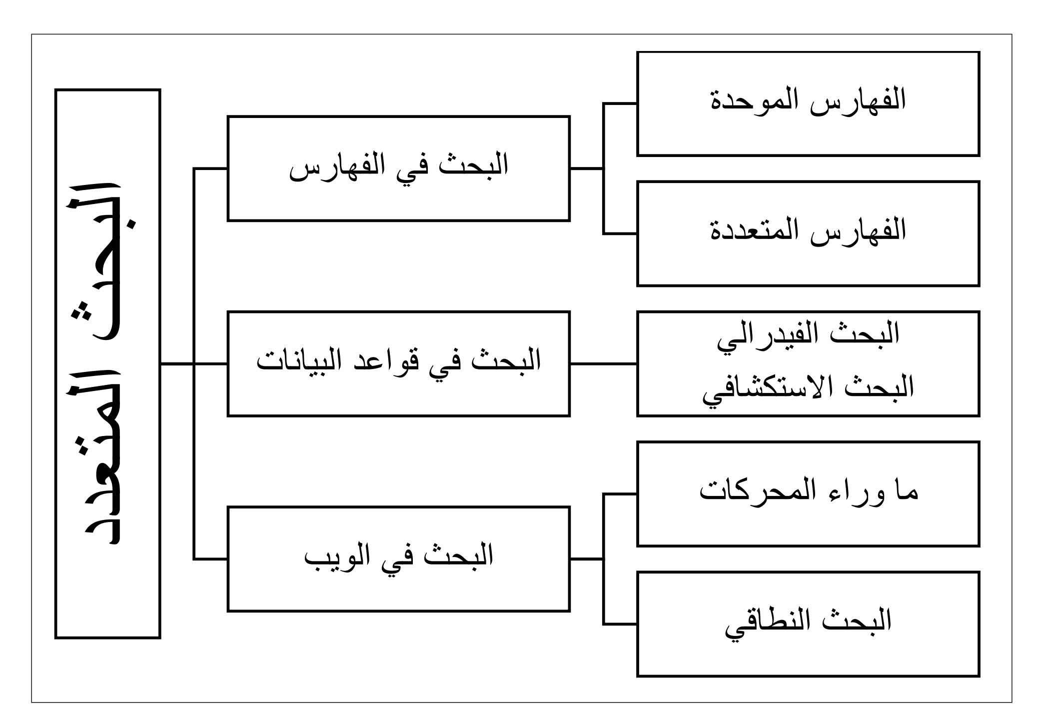
وقد صدر هذا المعيار SRW\U عن مكتبة الكونجرس الأمريكية ويعد أحد المعايير

الأساسية التي تراعيها المكتبة في تطبيقات نظم استرجاع المعلومات الببليوجرافية (Library of congress, 2008). ويساعد بروتوكول SRW\U على إجراء البحث المتعدد من خلال وكيل بحث يقوم بإجراء البحث في قواعد البيانات المتاحة على الويب واسترجاع النتائج بسلاسة دون الحاجة إلى استخدام بروتوكول Z39.50 الأكثر تعقيداً (Morgan,2004). فعند المقارنة بين بروتوكول SRW\U وبروتوكول Z39.50 نجد أن بروتوكول SRW\U أكثر سهولة في التطبيق ويؤدي نفس الوظيفة الدلالية لبروتوكول Z39.50 (Levan, 2003, Mohamed, 2004, 2015, 2016).

وإلى جانب التحديات التي سبق ذكرها فيما يتعلق بالبحث المتعدد، فإن دمج النتائج Results Merging التي يتم استرجاعها من قواعد البيانات المتعددة يُعد أيضاً من الأمور المهمة في هذا المجال. فعلى سبيل المثال أصبح أسلوب عرض النتائج مرتبة نموذجاً ومطلباً أساسياً متزايداً في بيئة الويب. فمن غير الطبيعي أن نتوقع حصول النتيجة التي جاءت في الترتيب رقم 1 من نظام استرجاع معين على نفس الترتيب عند دمج النتائج مع نتيجة أخرى حصلت على ترتيب رقم 1 من نظام آخر، وعادة ما يتم استخدام أساليب دمج البيانات Data Fusion كنموذج لدمج النتائج في البحث المتعدد بقواعد البيانات للحصول على أفضل قائمة نتائج مرتبة عند استخدام هذه الحلول. وقد اختبر خالد عبدالفتاح محمد (Mohamed,2004) ثلاث خوارزميات وبدائل دمجها وتدويرها، لإجراء الدمج والفرز للنتائج من ثلاث محركات بحث، وتوصل إلى أنه لا توجد خوارزمية دمج تحقق نتائج أفضل من باقي الخوارزميات وأنه لابد من الدمج بين أكثر من حل من الحلول المنطقية التي يتم تطبيقها على الهواء On the Fly عند دمج وترتيب النتائج المسترجعة من أكثر من محرك بحث لأغراض بناء ما وراء المحركات. وقد خصص مؤتمر TREC مساراً خاصاً لدمج وفرز النتائج لأغراض البحث المتعدد من المصادر غير المتجانسة Heterogenous وعرضها في قائمة موحدة (Voorhees & Hanman, 2000).

وعادة ما يتم تطبيق البحث في قواعد البيانات المتعددة من خلال موردي قواعد البيانات مثل Proquest, EBSCOHOST, DIALOG كما أن ما وراء محركات الويب

تـؤدي وظيفـة شـبيهة بعمليـات البحـث المتعـدد بقواعـد البيانـات والفهـارس، حيـث تسترجع النتائج مـن أكثـر مـن محـرك بحـث واحـد علـى الإنترنـت. ويوجـد ثلاثـة أنـواع أساسية للبحـث فـي المصـادر المتعـددة يوضحهـا الشـكل التالـي:



7.4.1 الفهارس

يشتمل هذا النوع على نمطين أساسيين هما:

- **النمـط الأول:** البحـث فـي الفهـارس الأخـرى ويعتمـد علـى اسـتخدام بروتوكـول Z39.50 لربـط فهـرس المكتبـة بفهـارس المكتبـات الأخـرى، مـا يُمكـن المسـتفيد مـن البحـث فـي تلـك الفهـارس عنـد الحاجـة.
- **النمـط الثانـي:** يستخدم في بناء الفهـارس الموحـدة والـذي يعتمـد أيضاً اسـتخدام أسـلوبين أساسـيين فـي البنـاء همـا (محمـد، 2011):

• الفهـارس الموحـدة المركزيـة Physical Union Catalogs والتـي تقـوم بتجميـع

بيانات كل الفهارس المستقلة في فهرس واحد مركزي يستخدم في عمليات البحث المجمع. ويعتمد هذا النوع على بروتوكول Z39.50 في تجميع التسجيلات من الفهارس المستقلة.

- الفهارس الموحدة التخيلية Viritual Union Catalogs التي يتم فيها بناء واجهة موحدة يمكن من خلالها البحث في كل الفهارس المستقلة دون الحاجة إلى تجميع الفهارس في قاعدة بيانات موحدة مع إجراء عمليات الدمج والفرز وفقاً لآليات وخوارزميات متنوعة. وتستخدم هذه الفهارس مزيجاً من بروتوكولات Z39.50 وبروتوكولات الروابط المفتوحة SRW/U.

7.4.2 البحث في قواعد البيانات المتعددة

يوجد أسلوبان أساسيان شائعان الآن لهذا النمط من أنماط البحث هما:

- البحث الفيدرالي Federated Search والذي يعتمد على نفس أسلوب الفهارس الموحدة التخيلية؛ حيث يستند إلى واجهة موحدة تقوم بتلقي استفسارات المستفيدين وإرسالها إلى قواعد البيانات المستقلة وتسترجع النتائج منها ثم تقوم بدمجها في قائمة موحدة وعرضها مرتبة للمستفيد وتتم عملية التجميع والفرز على الهواء On the fly.

- البحث الاستكشافى Discovery Search: ويعتمد هذا النمط على نفس أسلوب عمل الفهارس الموحدة المركزية؛ حيث يقوم بتجميع كل التسجيلات في قاعدة بيانات ميتاداتا موحدة تستخدم للبحث في قاعدة البيانات المركزية دفعة واحدة، بدلاً من إجراء البحث في قواعد البيانات المستقلة. من ثم فعملية التجميع والدمج تتم قبل إجراء البحث في مقابل إجراء التجميع والبحث على الهواء في البحث الفيدرالي.

1. الويب: يتم البحث في شبكة الويب بالاعتماد على آليات استكشاف مصادر

المعلومات المتاحة من خلال محركات البحث. بمعنى آخر أنه يستخدم إمكانيات محركات البحث في استكشاف شبكة الويب بالاعتماد على آليات عمل تلك المحركات والتي تستخدم أدوات مثل الزواحف Crawlers. وتوجد طريقتان أساسيتان يمكن من خلالهما استكشاف محركات البحث هما:

- ما وراء المحركات Meta Search Engines وهي عبارة عن أداة بحث تستطيع البحث في أكثر من محرك في نفس الوقت. تقوم تلك الأداة بتلقي استفسارات المستفيدين وإرسالها إلى محركات البحث المتعددة واستقبال النتائج من تلك المحركات وإجراء عمليات الدمج. بمعنى إنشاء قائمة نتائج موحدة وفرز تلك النتائج وفقاً لإحدى خوارزميات الفرز ثم عرض النتائج بأسلوب موحد للمستفيد على واجهة أداة البحث.
- البحث النطاقي للويب Web Scale Searching يعتمد هذا النمط على استخدام إمكانيات محركات البحث في إجراء استكشاف لقطاع موضوعي أو مجموعة محددة من القطاعات بقاعدة بيانات أو بمجموعة من قواعد البيانات أو المحركات أو نوعية معينة من المصادر سواء كانت نوعية معينة من الوثائق مثل الصور أو الملفات المسموعة أو الفيديو image, Video, youtube أو قطاعاً معيناً مثل الوثائق العلمية كما هو الحال Google Scholar, Pubmed, CiteseerX. وهو في هذه الحالة يشبه البوابات المتخصصة في قطاعات موضوعية معينة أو فئات معينة من الوثائق لكنه يركز البحث في نطاق محدد من الوثائق بمحركات البحث.

7.5 اختيار آلية البحث

اتضح من العرض السابق أنه توجد أدوات بحث متنوعة يمكن للمستفيد النهائي أن يستخدمها ويوظفها لإجراء عمليات البحث عن المعلومات. وتوجد العديد من العوامل التي يجب أن يراعيها المستفيد عند اختيار آلية البحث الملائمة. وسوف تركز المناقشة في هذا الجزء على اختيار آلية البحث بناءً على وظائفها وأداء نظام استرجاع المعلومات.

◄ 7.5.1 وظائف آليات الاسترجاع

تعمل آليات الاسترجاع المختلفة بأساليب متنوعة، ولكل آلية طريقة في الأداء تساعد المستفيد على تحقيق أهدافه من البحث بشرط استخدام الطريقة الملائمة في الموقف البحثي. فعلى سبيل المثال استخدام البتر يساعد على استرجاع الأشكال المختلفة للمصطلح والتي تتشابه معاً في أجزاء من هجائها وشكل كتابتها وتحمل معنى مشتركاً أو مرتبطاً. ويقوم البحث الغامض أو المجرد بالتعامل مع أخطاء الهجاء وبرامج التعرف الضوئي على الحروف في حالة المطابقة أو المضاهاة بين الشكلين المختلفين للمصطلح. لذلك فإن السؤال الأول الذي يجب أن يسأله المستفيد قبل إجراء البحث، وبعد تحديد سلة المصطلحات اللازمة للبحث، هو ما هي آلية البحث الملائمة لتحقيق الهدف من استرجاع المعلومات. وبمجرد الإجابة عن هذا السؤال يستطيع المستفيد تحديد الآلية الملائمة لطبيعة العبارة البحثية التي يرغب في البحث عنها.

◄ 7.6 أداء نظام استرجاع المعلومات

عادة ما يتم قياس أداء نظم استرجاع المعلومات بالاعتماد على مقاييس الاستدعاء والتحقيق، على الرغم من أن هذين المقياسين هما محل جدل دائم بين المتخصصين. وسوف يركز هذا القسم على الاستدعاء والتحقيق كمقياسين من مقاييس الأداء وسوف يترك الجدل الدائر حولهما للدراسات التي تناولت تقييم الأداء في نظم استرجاع المعلومات.

التحقيق Precision يتم حساب معدل الوثائق الصالحة المسترجعة إلى إجمالي عدد الوثائق المسترجعة من النظام؛ حيث يختبر هذا المقياس قدرة النظام على الفصل، بمعنى قدرته على عزل الوثائق غير الصالحة واسترجاع الوثائق الصالحة فقط. نفترض أنه تم استرجاع 100 وثيقة لاستفسار معين، وتم الحكم على 35 وثيقة فقط منها أنها صالحة، يكون معدل التحقيق في النظام هو 35 ٪.

التحقيق = عدد الوثائق الصالحة المسترجعة / إجمالي عدد الوثائق المسترجعة × 100

الاستدعاء Recall يتم حسابه بمعدل الوثائق الصالحة المسترجعة إلى إجمالي عدد الوثائق الصالحة في النظام بأكمله. ويختبر هذا المقياس القدرة الاسترجاعية Retrievability لنظام استرجاع المعلومات. نفترض أنه يوجد 100 وثيقة صالحة في النظام بأكمله في موضوع معين، عند إجراء البحث في النظام عن هذا الموضوع، تم استرجاع 45 وثيقة فقط من ثم يكون معدل الاستدعاء في هذا النظام 45 ٪.

الاستدعاء = عدد الوثائق الصالحة المسترجعة / إجمالي الوثائق الصالحة في النظام × 100

وعلى الرغم من أنه كلما ارتفعت النسبة التي يتم حسابها لأي من المقياسين، كان أداء النظام أفضل؛ إلا أنه من المستحيل الحصول على نسبة مرتفعة للمقياسين معاً وذلك لوجود علاقة عكسية بينهما، والتي تشير إلى أنه كلما ارتفعت نسبة التحقيق انخفض نسبة الاستدعاء والعكس. ويرجع ذلك إلى أن الجزء الأول من المعادلة في كل من المقياسين ثابت والاختلاف في الجزء الثاني.

وبالنظر إلى أداء نظم استرجاع المعلومات من حيث آليات الاسترجاع فإنه يمكن تقسيم تلك الآليات إلى:

- آليات تحسن التحقيق مثل استخدام المعامل البوليني AND والبحث بالوزن النسبي.
- آليات تحسن الاستدعاء مثل المعامل البوليني OR والبحث المجرد.

لذلك، فإن اختيار آلية البحث لابد أن تراعي مستوى الأداء الاسترجاعي الذي يرغب المستفيد في تحقيقه من العبارة البحثية، فإذا كان المستفيد يرغب في مستوى عال من التحقيق فعليه اختيار الآلية الملائمة لذلك الغرض والعكس.

7.6.1 آليات الاسترجاع لتحسين التحقيق

يساعد **المعامل البوليني AND** على تحسين مستوى التحقيق من خلال المزج بين مصطلحين في العبارة البحثية لتحديد مستوى الدقة اللازم في العلاقة بين المفاهيم

عند إجراء البحث. فعلى سبيل المثال إذا كان المستفيد يرغب في البحث عن المصطلحات الثلاثة: تسوية، النزاعات، الإقليمية، فإنه يمكنه الحصول على نتائج دقيقة من خلال استخدام المعامل البوليني AND في الربط بين المصطلحات الثلاثة. أما إذا تم استخدام مصطلحين فقط في العبارة البحثية واستبعاد الثالث، فإن عدد الوثائق المسترجعة سوف يرتفع وينخفض معه عدد الوثائق الصالحة وينخفض معه مستوى الدقة نظراً لعدم تقييد البحث باستخدام المصطلح الثالث.

المعامل البوليني NOT يساعد أيضاً على تحسين مستوى الدقة في النتائج المسترجعة من خلال حذف المصطلحات التي لا يرغب المستفيد في استرجاعها ضمن قائمة النتائج. نفترض أنه يوجد مستفيد يرغب في البحث عن وثائق تسوية النزاعات الإقليمية وليس الدولية، فإن المعامل البوليني NOT يجب أن يستخدم في هذه الحالة لتحقيق الغرض من العملية البحثية. ويمكن صياغة الاستراتيجية كالتالي: (تسوية AND نزاعات AND أقليمية) NOT دولية. ولاحظ استخدام الأقواس لتحديد الأولويات البحثية.

يساعد البحث بالحروف الحساسة على زيادة الدقة من خلال التمييز بين الحروف الرومانية. فكما أوضحنا من قبل، عند البحث عن العلامة التجارية Target أو محال Target يتطلب كتابة الحرف T الكبير أما عند الحاجة إلى البحث عن المصطلح target بمعنى هدف أو غاية، فإن المستفيد في هذه الحالة بحاجة إلى استخدام حرف الـ t الصغير. وإذا كان النظام لا يتيح إمكانية إجراء البحث بالحروف الحساسة، وهو الحال في الغالبية العظمى من النظم الحالية، بالتالي لن يكون أمام المستفيد أي خيار في التمييز بين الحروف. من ثم سيقوم النظام باسترجاع كل الوثائق التي تتناول المصطلح target,Target دون تمييز بين دلالة المصطلح في كل حالة، ما يؤثر في معدل دقة أداء نظام استرجاع المعلومات بصورة سلبية.

المعامل with الذي يستخدم في البحث بالتقارب يساعد أيضاً على تحسين مستوى الدقة في النتائج، نظراً إلى أنه يحدد الترتيب الذي يجب أن تظهر فيه المصطلحات في النتائج المسترجعة كما وردت في العبارة البحثية (الاستفسار). فعند البحث عن

المصطلحين information with technology لابد أن يسترجع النظام وثائق تتناول الموضوعات بنفس الترتيب، ويتم استبعاد أي وثائق تشتمل على أي مزيج مخالف للترتيب الوارد في الاستفسار مثل (technology information , information and technology) حيث إنها سوف تسترجع نتائج غير دقيقة بناء على الترتيب الذي حدده المستفيد في الاستفسار الأساسي.

كذلك الحال بالنسبة للمعامل n with فإنه يساعد على تحسين مستوى الدقة، حيث إنه يحدد عدد الكلمات التي تفصل بين المصطلحات المستخدمة في الاستفسار مع مراعاة الترتيب الوارد في صياغة الاستفسار وفقاً لعدد n من الكلمات التي يربط بينها المعامل.

- **ضبط المسافات + BOLD** يُعد أيضاً من آليات تحسين مستوى الدقة في الاسترجاع من خلال إعطاء وزن نسبي لكل مصطلح من المصطلحات المستخدمة في الاستفسار، ما يساعد المستفيد على التركيز على جانب من جوانب الموضوع بصورة أكبر والحصول على نتائج مطابقة لتوقعاته. فعلى سبيل المثال عند البحث عن موضوع (تسوية النزاعات الإقليمية) في محركات البحث فإنه يمكن إعطاء تركيز أكبر على أحد الجوانب من خلال وضع علامة (+) بجوار المصطلح وترك المصطلح الآخر من دون أي علامة مميزة (+تسوية + النزاعات الإقليمية). وتعني هذه العبارة البحثية أن المستفيد مهتم أكثر بموضوعي (تسوية) و (النزاعات) ويجب تسليط الضوء على هذين الجانبين عند إجراء البحث. من ثم فإن استخدام آليات الوزن النسبي للمصطلحات يساعد على تحقيق مستوى أكبر من الدقة في النتائج المسترجعة وفقاً لنقاط التركيز التي يراها المستفيد.

- **البحث الحقلي Field Searching** يساعد على تحقيق الدقة في البحث من خلال تقييد البحث في حقول معينة؛ حيث إن كل حقل من الحقول المستخدمة في التمثيل يمثل محدداً معيناً في الوثيقة. فإن كان البحث عن وثائق لمؤلف معين فإن المستفيد هنا بحاجة إلى إجراء البحث عن هذا المؤلف باسمه مع تقييد البحث في حقل المؤلف. من ثم يحصل على نتائج

أكثر دقة عند تقييد البحث في حقل المؤلف من تركها عامة في كل الحقول؛ حيث إنه من الممكن أن يرد اسم هذا المؤلف في حقول أخرى في الوثيقة لا تعكس دوره كمؤلف.

والخلاصة أن المعاملات البولينية AND, NOT والبحث بالحروف الحساسة والمعامل with المستخدم في البحث بالتقارب والمعامل n with والبحث الحقلي والبحث بوزن المصطلحات كلها آليات تستخدم في تحسين مستوى الدقة في النتائج المسترجعة.

7.6.2 آليات الاسترجاع لتحسين الاستدعاء

في بعض الأحيان قد يحتاج المستفيد إلى توسيع نطاق البحث للحصول على عدد أكبر من النتائج وتغطية كافة عناصر الموضوع الذي يتناوله بمفاهيمه المتنوعة وسلة المصطلحات التي حددها. وتوجد مجموعة من الآليات التي تساعد على توسيع نطاق البحث تشمل ما يلي:

- المعامل OR: ويستخدم المعامل OR لتوسيع نطاق البحث، حيث إنه يستخدم لاسترجاع أي وثيقة يظهر بها أي مصطلح من المصطلحات المربوطة بالمعامل OR، بالتالي يرتفع عدد النتائج المسترجعة ويرتفع معه معدل الاستدعاء. فعلي سبيل المثال عند البحث عن الانتخابات أو التصويت فإن النظام سوف يسترجع أي وثيقة يرد بها أي من المصطلحين إلى جانب استرجاع الوثائق التي يرد بها المصطلحان معاً. من ثم فإن المعامل البوليني OR لا يضع أي قيود في عملية البحث تؤدي إلى تضيق النطاق مقارنة بالمعاملين الآخرين AND /NOT. وتجدر الإشارة إلى أنه كلما قلّت القيود أو المحددات، ارتفع عدد الوثائق المسترجعة وارتفع معها الاستدعاء.
- البتر يساعد على توسيع نطاق البحث من خلال استخدام الجزء المشترك من المصطلح في الاستفسار (مثل جذر الكلمة Word Stem)، واسترجاع كل الأشكال المختلفة في قائمة النتائج. فعلى سبيل المثال عند إجراء بحث

بالبتر عن المصطلح (*ejournel) فإن النظام سوف يسترجع كل الوثائق التي تشتمل على المصطلحات (ejournels, ejournel, ejournalist, ejournalism, etc.) أو غيرها من المصطلحات التي تبدأ بالجزء ejournal. ومن الواضح أن معدل الاستدعاء لعملية البتر في هذه الحالة سوف يرتفع نتيجة لتوسيع نطاق البحث، ويسترجع وثائق أكبر من حالة عدم البتر التي سوف تسترجع الوثائق التي تضمنت سلسلة الحروف الواردة في الاستفسار فقط.

- معامل التقارب near يساعد أيضاً على توسيع نطاق البحث، حيث يسمح للنظام باسترجاع المصطلحات التي يتم ربطها بالمعامل near بصرف النظر عن ترتيبها في الوثائق المسترجعة. من ثم فإن استخدام المعامل near في الاستفسارات مثل information near technology سوف يسترجع وثائق تتناول information technology ووثائق من technology information ما يساعد على رفع معدلات الاستدعاء في النتائج المسترجعة. ويعمل المعامل near n بنفس الطريقة التي يعمل بها المعامل near مع تحديد عدد الكلمات التي يجب أن ترد بين المصطلحين اللذين تم ربطهما معاً بالمعامل near.

- البحث المجرد يستخدم أيضاً وسيلة من وسائل توسيع نطاق البحث من خلال تحديد وتصحيح الأخطاء التي تحدث نتيجة أخطاء الهجاء أو أدوات التعرف الضوئي إلى الحروف وغيرها. فإذا كانت الوثيقة تتناول موضوع cellular phone والمستفيد أخطأ في كتابة المصطلح وكتبه celluler؛ فإن النظام سيظل قادراً على استرجاع الوثيقة في حال استخدام إمكانيات البحث المجرد، من ثم فإن النظام في هذه الحالة يساعد على رفع معدلات الاستدعاء.

- تعديل الاستفسار: توسيع الاستفسار يهدف إلى استرجاع عدد أكبر من الوثائق الصالحة من خلال تعديل الاستفسارات بناء على استخدام دفعة من النتائج الأولية في تحسين كفاءة الاستدعاء. ويمكن أن يتم تكرار عمليات التعديل وتوسيع الاستفسارات حتى يتم الحصول على العدد الكافي من الوثائق الصالحة، فعلى سبيل المثال نفترض أنه عند البحث بمصطلح غير متداول

كثيراً مثل vector space model قام النظام باسترجاع 5 وثائق فقط، واستنبط النظام من هذه الوثائق أن اسم (Salton) كان شائعاً في هذه الوثائق الخمس. من ثم يمكن استخدام الاسم في إجراء البحث في النظام في جولة ثانية مع المصطلح العام مثل retrieval، بالتالي يستطيع النظام أن يسترجع عدداً آخر من الوثائق في الدفعة الثانية تضاف إلى الدفعه الأولى لتحسين مستوى الاستدعاء.

- البحث في المصادر المتعددة يُعد أيضاً من آليات تحسين مستوى الاستدعاء، بسبب استخدام أكثر من قاعدة بيانات واحدة في البحث، ما يعطي الفرصة لاسترجاع عدد أكبر من الوثائق الصالحة من التي يتم استرجاعها من قاعدة بيانات واحدة.

من ثم يمكن القول إن المعامل البوليني OR والبتر ومعاملات البحث بالتقارب near, n near والبحث الغامض أو المجرد وآليات توسيع وتعديل الاستفسارات والبحث في قواعد البيانات المتعددة كلها آليات تساعد على توسيع نطاق البحث بطريقة أو بأخرى. وعلى الرغم من أنه ليس شرطاً أن تحقق زيادة عدد النتائج المسترجعة مستوى مرتفعاً من الاستدعاء؛ لأنها يجب أن تكون نتائج صالحة؛ إلا أنها ترتفع معها احتمالات زيادة معدلات الاستدعاء لأي استفسار. بالتالي فإن المستفيد يجب أن يكون على وعي كامل كيف يؤثر كل أسلوب من أساليب البحث في معدلات الاستدعاء والدقة في عمليات البحث حتى يستطيع المستفيد اتخاذ القرار المناسب واستخدام آلية البحث الصحيحة التي تتناسب مع احتياجاته.

7.7 تمثيل الاستفسارات
query representation

يتم التعبير عن الاحتياجات المعلوماتية لفظياً باستخدام المصطلحات الملائمة قبل إجراء عملية البحث ويطلق على الاحتياجات المعلوماتية التي يتم صياغتها في صورة مجموعة من المصطلحات التي يتم الربط بينها (طلبات البحث والاسترجاع باستخدام اللغة الطبيعية). ويتم تحويل طلب البحث إلى استفسار باستخدام

إمكانيات نظم استرجاع المعلومات مثل بنية الاستفسار Query Syntax وتقنيات الاسترجاع Retrieval Techniques والمصطلحات المضبوطة في حال استخدامها. ويطلق على عملية تحويل الاحتياجات المعلوماتية إلى عبارة بحثية مصطلح (تمثيل الاستفسار)، والذي يُعد أهم العناصر المؤثرة في عملية البحث وأداء نظم استرجاع المعلومات (Sparck, 2000).

7.7.1 خطوات تمثيل الاستفسارات

تُعد عملية تمثيل الاستفسارات إجراءً فكرياً يتضمن من الخطوات التالية:

1. إجراء تحليل مفاهيمي لطلب البحث من خلال تحليله إلى مجموعة من المفاهيم أو الأوجه.
2. إعداد سلة المصطلحات الخاصة بكل مفهوم والتي تشمل المترادفات والمصطلحات الأوسع والأضيق.
3. ترجمة المصطلحات إلى لغة النظام سواء كانت اللغة الطبيعية أو المضبوطة ولكن بصفة عامة يفضل استخدام القواميس والمكانز المتخصصة والعامة عند ترجمة المفاهيم إلى مصطلحات بحثية.
4. إعداد استراتيجية البحث والتي تشمل الربط بين المصطلحات والمفاهيم باستخدام المعامل البوليني OR مع المترادفات، والمعامل البوليني AND للربط بين المفاهيم، والمعامل NOT لاستبعاد أحد أوجه المفاهيم غير المطلوبة في الاستفسار.
5. تطبيق آليات البحث والاسترجاع الأخرى مثل البحث المجرد أو البحث الحقلي.. إلخ في حالة الحاجة إليها.

وعلى الرغم من أن هذه الخطوات ما هي إلا مجرد تعليمات لممارسات شائعة ومقترحة؛ إلا أنها تتضمن جوهر عملية تمثيل الاستفسارات. ومن الممكن أن تكون

هناك مجموعة من الاختلافات في الممارسة الفعلية، ويتم فيما يلي مناقشة عملية تمثيل الاستفسارات خطوة بخطوة مع مراعاة دورها الرئيس في عمليات استرجاع المعلومات.

7.7.1.1 تحليل المفاهيم
concept analysis

يتم في المرحلة الأولى من تمثيل الاستفسارات تحليل طلب البحث إلى مجموعة المفاهيم الأساسية أو الأوجه Facts، فعلى سبيل المثال إذا كان طلب المعلومات هو الحصول على الوثائق التي تتناول الموضوع التالي:

تسوية الصراعات في الشرق الأوسط.

فبتحليل الطلب السابق نجد أنه يشتمل على ثلاثة مفاهيم مختلفة كما يوضحها الجدول التالي:

المفهوم (1)	المفهوم (2)	المفهوم (3)
تسوية	صراعات	الشرق الأوسط

جدول 6.2 تحليل مفاهيم طلب البحث

في هذه الحالة من الممكن أن تكون المصطلحات المستخدمة في عملية البحث هي نفسها التي تعبر عن المفاهيم، إلا أن هناك حالات تظهر فيها اختلافات ما بين المفاهيم والمصطلحات، ولا توجد مضاهاة كاملة بين المصطلحات والمفاهيم. فعلى سبيل المثال قد يكون طلب المستفيد مشتملاً على الحاجة إلى معلومات عن الأتوبيسات buses ومترو الأنفاق subways إلا أن تحليل الطلب قد يوضح أن المستفيد بحاجة إلى استخدام مصطلح (المواصلات العامة) public transportation في البحث بدلاً من الأتوبيسات ومترو الأنفاق في تمثيل الاستفسار، إضافة إلى ذلك يجب استخدام الأعلام والمسميات الاصطلاحية في جمل اسمية Noun Phrases،

في تمثيل المفاهيم. ويتم تمثيل الأفعال التي ترد في الطلبات باستخدام معاملات الربط البوليني، أما الأجزاء الأخرى من الطلب مثل الحروف والكلمات الوظيفية فلا يتم استخدامها في تمثيل المفاهيم التي ترد في طلبات المستفيدين. ومن ثم فإن تحليل المفاهيم يركز على الأسماء الاصطلاحية والجمل الاسمية التي ترد في طلبات المستفيدين ويقوم بتحويل هذه المفاهيم إلى مصطلحات.

7.7.1.2 تنوع (أشكال) المصطلحات
Term variations

تتنوع المصطلحات في معظم الحالات ما بين مترادفات، مصطلحات أوسع، مصطلحات أضيق وغيرها من الأشكال. والغرض الأساسي من عملية تحديد المصطلحات هو تجميع كل الأشكال المختلفة للمصطلحات الدالة على المفاهيم التي تم تحديدها في الخطوة السابقة؛ بحيث يتم تمثيل المفهوم بصورة شاملة ويوضح الجدول 6.2 الأشكال المحتملة لمفهوم تسوية الصراع في الشرق الأوسط مع إضافة أن المطلوب هو وثائق من الويب والذي يمكن التعبير عنه كما يلي.

جدول (7.1) تقسيم المفاهيم وبناء سلة المصطلحات

Concept 1	Concept 2	Concept 3
Settlelement	Controversy	Middle East
Adjustement Compromise Equalization Normalization Conciliation	Depate Dispate Conflect	Meddle East MENA Arab Countries And Israel Iran And Israel Arab Countries And Iran

ويتضح من الجدول السابق أنه ليس شرطاً أن تكون كل بدائل المصطلحات وأشكالها المختلفة مستخدمة ومعروفة من جانب المستفيدين، وأن المستفيد في الغالب يركز على المصطلحات الشهيرة والمختصرات، فعلى سبيل المثال نلاحظ أن المفهوم الأول لم يشتمل على المصطلح intercession والذي يشير إلى الوساطة، وأن قرار إدراج مصطلح من عدمه يعتمد على معايير ذاتية مثل توقعات المستفيد والاستدعاء المتوقع من جانب المستفيد ومدى تأقلمه مع الموضوع ومصطلحاته. فعلى الرغم من أن إدراج كل المصطلحات وأشكالها المختلفة وبدائلها المتنوعة في الاستفسار النهائي قد يؤثر في عملية البحث، إلا أن ذلك سوف يساعد المستفيد بعد الجولة الأولى من الاستفسار على تحديد المصطلحات القابلة للبحث بدقة. بالتالي يجب أن يفهم المستفيد أن عملية البحث تتم بأسلوب الاستفسار والبحث والتفتيش Quering, Searching, Snooping وأن عملية البحث هي عملية مستمرة تتم على جولات متعددة حتى يصل المستفيد إلى أفضل النتائج.

ويساعد هذا الإجراء على تحديد كل الأشكال والبدائل المختلفة للمصطلح، والذي يتطلب الرجوع إلى قائمة المصطلحات المضبوطة والمعاجم اللغوية والمتخصصة والأنطولوجيات وقوائم الكلمات والتقسيمات إلى فئات.. الخ.

◄ 7.7.1.3 تحويل المصطلحات

Terms conversion

عند استخدام النظام لقائمة مصطلحات مضبوطة في عمليات التمثيل بنظام استرجاع المعلومات، فإنه يجب تحويل المصطلحات التي يتم التعبير عنها باللغة الطبيعية إلى نظام المصطلحات المستخدم بالنظام. أما في حالة استخدام اللغة الطبيعية في التعبير عن المصطلحات، فإنه يجب الالتزام باللغة الطبيعية في تعبير عن المصطلحات مع إثراء مصطلحات الاستفسار من خلال الأدوات المساعدة مثل القوائم المضبوطة والقواميس. وتتطلب عملية تحويل المصطلحات أن يكون المستفيد على دراية ووعي بكيفية توظيف اللغة المضبوطة المستخدمة بالنظام، ويمكنه استخدام أي من الأساليب التالية:

I. المطابقة الكاملة Exact Equivalent

المطابقة الكاملة تعني استخدام المصطلح المخصص والمطابق بالكامل للمفهوم الذي يسعى المستفيد إلى البحث عنه من قائمة المصطلحات المضبوطة. ويُعد هذا الأسلوب أسهل أساليب تحويل المصطلحات، فعلى سبيل المثال إذا كان المستفيد يبحث عن الشرق الأوسط فالمطابقة التامة هنا تعني استخدام مصطلح موازٍ تماماً للمفهوم دون التوسيع أو التضييق.

II. استخدام المترادفات والمصطلحات المرتبطة Synonyans or Related Terms

يهتم هذا التوجه بالاعتماد على قوائم المصطلحات المضبوطة لاشتقاق المترادفات والمصطلحات المرتبطة بالمفهوم، بالتالي لابد أن يبذل المستفيد جهداً إضافياً في عملية اختيار هذه النوعية من المصطلحات من قائمة المصطلحات المضبوطة، والتي تُعد قريبة في المعنى من المصطلح الذي يبحث عنه المستفيد.

III. استخدام المصطلح الأوسع Broader Terms

إذا لم توجد مصطلحات مساوية أو مترادفات للمفهوم الذي يبحث عنه المستفيد يجب استخدام المصطلح الأوسع في الدلالة على المفهوم، كما يجب استخدام المصطلح الأوسع في الحالات التي قد يتأثر فيها البحث سلباً عند استخدام المصطلح المخصص في عملية تحويل المصطلحات.

IV. استخدام المصطلح الأضيق Narrower Terms

في بعض الأحيان قد يكون للمفهوم الذي يبحث عنه المستفيد مصطلحات أضيق في الدلالة على المعنى ولا يوجد له مصطلحات مساوية أو مرادفات أو مصطلحات أوسع منه. في هذه الحالة يضطر المستفيد إلى استخدام المصطلحات الأضيق في الدلالة على المفهوم، من ثم يتم تقسيم المفهوم الذي يبحث عنه المستفيد إلى نطاقات أو مجموعة من المصطلحات الأضيق.

٧. استخدام الأسماء

أحياناً قد يبحث المستفيد عن أسماء مثل أسماء الشركات أو الأشخاص أو المنتجات، أو الأماكن.. إلخ أو غيرها من الأسماء الجديدة التي لا يوجد لها بدائل موازية بقوائم المصطلحات المستخدمة في النظام. وفي هذه الحالة لابد من استحداث مصطلح يُطلق عليه مُحدد Identifier لإجراء عملية التحويل الاصطلاحي. ومن الوارد جداً أن يكون المصطلح الجديد هو المحدد الذي تم تجهيزه لأغراض التحويل. وباستثناء عملية استخدام المصطلح المساوي، فإن كل أساليب التحويل الأخرى تتطلب عملية تفسير للمفاهيم لأغراض التحويل. وتؤثر دقة عملية تفسير المفاهيم في دقة المصطلحات التي يتم تحويلها للتعبير عن المفاهيم التي يرغب المستفيد في البحث عنها.

7.8 تطبيق المعاملات البولينية

Application of bolean operators

نفترض أن المصطلحات التي تم تجميعها في جدول (7.2) للدلالة على المفاهيم الثلاثة التي يبحث المستفيد عنها تمثل الأشكال الصحيحة للمصطلحات الملائمة، من ثم فالخطوة التالية هي تطبيق المعاملات البولينية في الربط بين المصطلحات المختلفة الدالة على المفاهيم الثلاثة السابقة. وعلى الرغم من وجود بعض الاختلافات في التطبيق توجد قاعدتان أساسيتان لتطبيق المعاملات البولينية:

1. ربط كل المصطلحات الدالة على نفس المفهوم والمصطلحات التي تنتمي إلى سلة مجموعة واحدة باستخدام المعامل OR
2. استخدام المعامل AND للربط بين المفاهيم المختلفة بمعنى الربط بين كل المجموعات، بحيث يمثل كل منها مفهوماً مختلفاً باستخدام المعامل AND وفي بعض الأحيان القليلة والاستثنائية استخدام المعامل NOT.

يوضح الجدول 7.2 هذه العملية كمثال للمفاهيم التي تم تجميع المصطلحات الدالة عليها في جدول 7.1.

Group 1	Group 2	Group 3
Settlelement OR Adjustement OR Compromise OR Equalization OR Normalization OR conciliation	Controversy OR Depate OR Dispate OR Conflect	Middle east OR MENA OR Arab Countries) AND Israel NOT (Iran
Group (1) AND	Group (2) AND	Group (3)

ويتضح من الجدول السابق أمران مهمان هما:

عدد المصطلحات التي تم استخدامها للدلالة على المفهوم الواحد والتي يُستخدم معها المعامل OR أو NOT ـ كما هو الحال في المفهوم الثالث الذي تم استخدام NOT معه لاستبعاد إيران من العبارة البحثية – يزداد كلما اتسع المصطلح وتعددت جوانبه. وهنا يرد سؤال مهم: هل هذه العملية لانهائية، بمعنى هل يجب استخدام كل المترادفات والمصطلحات المرتبطة والأوسع والأضيق والمساوية للدلالة، لبناء سلة المصطلحات الدالة على المفهوم؟

الإجابة بالطبع تتوقف على حجم النتائج التي يرغب المستفيد في الحصول عليها، إضافة إلى طبيعة تمثيل تلك النتائج بقاعدة البيانات، مع مراعاة أنه كلما ازداد عدد المصطلحات التي يتم ربطها باستخدام المعامل OR، ازداد عدد النتائج المسترجعة. وعلى الجانب الآخر كلما انخفض عدد المصطلحات التي يتم ربطها باستخدام المعامل OR، انخفض عدد النتائج المسترجعة الدالة على المفهوم أو المجموعة الواحدة. وفي حالة زيادة عدد المصطلحات على الحدود المقبولة (مصطلحان بحد أدنى وخمسة مصطلحات بحد أقصى لكل مجموعة)، يجب على المستفيد أن

يضع كل المصطلحات ويرتبها من حيث الأولوية والأهمية بالنسبة إليه، وأن يختار من بينها الأكثر دلالة على المفهوم الذي يرغب في البحث عنه، وأن يربط بينها باستخدام المعامل OR. ومن الواضح أن الشكل السابق لم يوضح عدد المصطلحات المستخدمة في الدلالة على كل مفهوم، حيث تم شرح المفهوم الخاص باستخدام المعامل OR، لأن قرار تحديد المصطلحات وأهميتها وأولويات البحث، قرار ذاتي يتعلق باحتياجات المستفيد ومدى عمقها ومدى أهمية كل مصطلح بالنسبه له.

الأمر الثاني الذي يجب توضيحه فيما يتعلق بالجدول 7.2 هو استخدام الأقواس، فعند مناقشة البحث البوليني سابقاً تمت الإشارة إلى عملية الترتيب في البحث البوليني المركب Combound Boolean Search، فالجدول 7.2 يمثل هذا النموذج من البحث الذي يتطلب استخدام الأقواس لتحديد الترتيب في عملية البحث المنطقي.

وفي حالة عدم استخدام الأقواس فإن المصطلح الأول في المفهوم الثالث (Middle East) عندما يتم ربطه أولاً بقائمة النتائج الخاصة بالمصطلح الأخير Conflect الخاص بالمفهوم الثاني Controversy سوف يؤثر في دقة النتائج التي يرغب المستفيد في الوصول إليها. لذلك لابد من استخدام الأقواس في العبارة البحثية لتحديد الترتيب وأولوية البحث عن المصطلحات في إطار علاقاتها بطلب المستفيد، بالتالي يتم البحث في المجموعة بالكامل ثم تحديد عدد النتائج المسترجعة لكل مجموعة وربطه بالمجموعة السابقة.

وتجدر الإشارة إلى أن البحث البوليني أثبت جدارته كأساس لعمليات البحث في معظم أنظمة استرجاع المعلومات؛ حيث إن المنطق البوليني هو المنطق الحاكم لعملية تمثيل استفسارات المستفيدين في معظم، إن لم يكن كل، حالات استرجاع المعلومات. إلا إذا كان المستفيد يحتاج إلى البحث عن مصطلح واحد فقط منفرد لا توجد له أي علاقات بمصطلحات أخرى، وهي عملية نادرة الحدوث. مع العلم أن عملية البحث البوليني تبدو أكثر تعقيداً من النموذج الموضح هنا وسوف يتم مناقشتها بالتفصيل في الفصل التالي الذي يتناول نماذج استرجاع المعلومات. وفي حالة عدم استخدام الأقواس لتجميع المصطلحات وتحديد أولوياتها وعلاقاتها؛ فإن

النتائج سوف تتأثر وقد يسترجع النظام العديد من الوثائق غير المرتبطة باحتياجات المستفيد، وذلك على افتراض أنه يتم استبعاد كل الأقواس من العبارة البحثية الموضحة في الشكل 7.2 ويتم الاحتفاظ بكل المصطلحات كما هي موضحة في الجدول 7.2 بنفس الترتيب.

7.9 استخدام آليات استرجاع أخرى

توجد العديد من الأساليب الأخرى التي يمكن أن يستخدمها المستفيد لتمثيل الاستفسار بدقة ووضوح. فعلى سبيل المثال يجب على المستفيد أن يراعي الاعتبارات التالية عندما يتعامل مع أي مفهوم:

- هل هناك حاجة إلى استخدام البحث بالحروف الحساسة في التفرقة بين المشترك اللفظي للمصطلح.
- هل توجد حاجة إلى استخدام معاملات التقارب with or near لتمثيل المصطلحات المركبة من كلميتن.
- هل يتم تحديد عملية البحث في حقول معينة مثل العنوان أو الكلمات المفتاحية.
- هل يدعم النظام المستخدم في البحث عملية البحث الغامض (المجرد).
- هل يمكن تحديد وزن نسبي للمصطلحات التي يتم البحث عنها لكل مفهوم.
- هل يوجد آلية لدعم الصلاحية الراجعة في النظام أو توجد آليات يدوية لتوسيع الاستفسار.
- هل يجب البحث في أكثر من قاعدة بيانات سواء بصورة مستقلة أو مجمعة.

سبق وأشرنا إلى أنه ليست كل نظم استرجاع المعلومات تدعم كل الآليات التي تمت مناقشاتها في هذا الجزء؛ لذلك فإن هذه القائمة من الأساليب وآليات البحث، تُعد قائمة مراجعة واختيار chick list أكثر منها، قائمة إجراءات must do list يتم استخدامها في عملية البحث.

وكما هو الحال في عملية تمثيل المعلومات فإن عملية تمثيل الاستفسارات أيضاً عملية صعبة معقدة. وعلى الرغم من تلخيص هذه الخطوات الخمس للتعبير عن الخطوات الرئيسة لتمثيل الاستفسارات، إلا أن الممارسة الفعلية من الممكن ألا تتضمن كل هذه الخطوات السابقة، وليس شرطاً أن يتم تطبيقها بنفس الترتيب ويتوقف الأمر على مدى خبرة المستفيد في التعامل مع نظم استرجاع المعلومات، حيث يتمكن المستفيد الخبير من دمج بعض الخطوات، بينما يحتاج المستفيد المبتدئ إلى تفاصيل أكثر، وقد لا يستطيع إجراء أي دمج للعمليات.

وبصفة عامة فإن الخطوة الأولى في عملية تمثيل الاستفسار تتعامل مع إعراب / الطلب Request Parsing أي تحليل الطلب إلى مفاهيم. وتتعامل الخطوتان الثانية والثالثة مع عملية ترجمة الاستفسار إلى مصطلحات، وتركز الخطوتان الرابعة والخامسة على تطبيق آليات مختلفة لإمكانيات نظام استرجاع المعلومات. ونظراً لأن كل مستفيد وكل طالب بحث، وكل نظام استرجاع معلومات كل منهم له ملامحه وسماته الخاصة؛ فإن عملية تمثيل الاستفسارات لابد أن تعكس هذه الظاهرة من خلال مراعاة هذه السمات المتنوعة.

7.10 صعوبات تمثيل الاستفسارات

تُعد عملية تمثيل الاستفسارات، كما أوضحت المناقشة السابقة، عملية فكرية وليست عملية آلية؛ حيث إنها تتطلب تفكيراً وتحليلاً وإصدار أحكام. وتوجد العديد من الصعوبات التي تواجه تلك العملية الفكرية هي:

ا. تحليل المفاهيم

يمثل تحليل المفاهيم الصعوبة الأولى في تمثيل الاستفسار؛ حيث يجب أن يكون لدى المستفيد المعرفة والخبرة والمهارة الكافية لتحديد والتعبير عن المفاهيم التي يتضمنها طلب البحث، وعدم الدقة في تحليل المفاهيم من أهم الظواهر السلبية التي تحدث في عملية البحث واسترجاع المعلومات.

II. اللغة

تعد صعوبة تمثيل اللغة هي الصعوبة الثانية في تمثيل الاستفسار، حيث إن اللغة الطبيعية لغة غنية، مرنة، واضحة إلا أنها أحياناً ما تكون غامضة. أما اللغة المضبوطة فهي صارمة اصطناعية، ومن الصعب صيانتها وتطويرها ومع ذلك يجب تمثيل مصطلحات الاستفسار بدقة باستخدام أي من اللغتين أو كليهما معاً. وقد تؤثر عملية التحويل وتؤدي إلى اختلافات في التمثيل، ما يؤثر في أداء نظام الاسترجاع. كما أن استخدام اللغة المضبوطة يزيد من الصعوبات من جانب المستفيد الذي يحتاج إلى وقت وجهد لكي يتأقلم ويتدرب عليها، وعلى الجانب الآخر فإن استخدام اللغة الطبيعية أيضاً له عيوبه التي تمت مناقشتها بالتفصيل في الفصل الرابع.

III. آلية الاسترجاع

يعد تطبيق آلية البحث والاسترجاع أحد الصعوبات التي قد تواجه عملية تمثيل الاستفسار، حيث إن كل نظام استرجاع معلومات له مواصفاته وآلية تطبيقه، بصرف النظر عن آلية الاسترجاع، حيث إن علامة (+) في بعض محركات بحث الإنترنت تستخدم بدلاً من المعامل البوليني AND وتستخدم في بعض النظم الأخرى لوزن المصطلحات، بمعني أنها تستخدم كعلامة للدلالة على أهمية المصطلح، من ثم فإن تَمكن المستفيد من تلك الآليات يحتاج أيضاً إلى وقت وتدريب وممارسة.

هذه الصعوبات قد تؤدي إلى مشكلات في تمثيل الاستفسارات، ما يؤثر في تحقيق المضاهاة ما بين تمثيل المعلومات وتمثيل الاستفسارات. والتغلب على تلك المشكلات يمكن من الناحية العملية من خلال تدريب المستفيد وتأهيله إلى جانب العمل على الجانب الآخر المتمثل في تطوير البحوث في مجال التمثيل الآلي للاستفسارات.

7.11 التمثيل الآلي للاستفسارات
Automatic Query Representation

يُعد هذا التوجه من المتطلبات التي تسعى النظم إلى تحقيقها، وهذه الطريقة تشبه غيرها من الطرق الآلية مثل التكشيف الآلي وغيرها من الطرق الآلية لمعالجة النصوص التي تعتمد على آليات مثل تردد المصطلحات، التقارب، وموقع المصطلح. وفي بعض الأحيان يتم تطبيق خوارزميات قائمة على نظرية الاحتمالات أو النماذج اللغوية أو آليات الذكاء الاصطناعي. وعلى عكس التكشيف الآلي الذي يشتمل على أنشطة آلية وفكرية؛ فإن تمثيل الاستفسارات يشتمل على مكون فكري فقط. ونظراً لأن الحاسبات مازالت لا تستطيع التفكير مثل الإنسان، فإنه مازال من الصعب التنبؤ أو تخيل الصعوبات التي تواجه العملية الفكرية المتعلقة بتمثيل الاستفسارات. وقد حظي هذا التوجه باهتمام كبير خلال المراحل الأولى لميكنة نظم استرجاع المعلومات، كما حظي باهتمام في مؤتمر استرجاع النصوص[1] TREC . وقد أشارت المرحلة الأولى من مؤتمر في نسخته TREC 1,2 إلى أن الاستفسارات المهيكلة آلياً تعمل بنفس كفاءة وقدرة الاستفسارات المهيكلة يدوياً في استرجاع المعلومات، وفي بعض الأحيان تؤدي بكفاءة أعلى من الاستفسارات اليدوية. وقد أشار سبارك جونز (Spark Jones,1995) إلى أنه لا توجد أي ميزة إضافية للاستفسارات اليدوية، وقد جرت بعض الدراسات في النسخة TREC 3,4 للمقارنة بين الاستفسارات القصيرة Short Queries وكان التوجه في النسخة TREC 5,6 هو المقارنة بين بناء الاستفسارات الطويلة بالطرق اليدوية والآلية واختبار كفاءة النظم عند التعامل مع كل منهما والمقارنة بينهما (Spark, Jones,2000).

وعلى الرغم من أن دراسات TREC ليست شاملة لكل عناصر الموضوع؛ إلا أنها أثارت قضية التوجه الآلي نحو بناء الاستفسارات والموقف الحالي للدراسات في هذا الاتجاه ويحتاج هذا الموضوع إلى دراسات مستقبلية لتحسين كفاءة الطرق الآلية لتمثيل الاستفسارات.

---
(1) TREC – Text Retrieval Conference

المصادر:

- محمـد، خالـد عبدالفتـاح (2011). الفهـرس الموحـد للمكتبـات الجامعيـة المصريـة، مجلـة الفهرسـت، ع 36، ص ص 105ـ29.
- Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. Information Processing & Management, 31(3), 431-448.
- Davis, Charles H. (1997.). From document retrieval to web browsing: some universal concerns. Journal of information, communication, and library science, 3 (3), 3-10.
- Gauch, S., Wang, J., & Rachakonda, S. M. (1999). A corpus analysis approach for automatic query expansion and its extension to multiple databases. ACM Transactions on Information Systems (TOIS), 17(3), 250-269.
- Grossman, D. A., & Frieder, O. (1998). Information retrieval: Algorithms and heuristics (Vol. 15). Springer Science & Business Media.
- Jones, K. S. (1995). Reflections on TREC. Inf. Process. Manage., 31(3), 291-314.
- Jones, K. S. (2000). Further reflections on TREC. Information Processing & Management, 36(1), 37-85.
- Kowalski, G. (1997). Information Retrieval System: Theory and Application.
- Levan, ralph. (2003).Z39.50 as a web service. Retrived December 1,2008, from staff.oclc.org/-levan/docs/srw-niso20030430.ppt
- Library of congress. (2008). SRU:Search/retrival via url.Retrieved December 1,2008, from www. Loc.gov/standards/sru
- Michael, J. J., & Hinnebusch, M. (1995). From A to Z39. 50: A networking primer. London: Mecklermedia,| c1994.
- Morgan, E. L. (2004). An introduction to the Search/Retrieve URL service (SRU). Ariadne, (40).
- Resnick, M. L., & Vaughan, M. W. (2006). Best Interface and future visions for search user interfaces. Journal of the American Society for Information Science and Technology, 57(6), 781-787.
- Salton, G. (1970). Automatic text analysis: science,168335-343.

- Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley.
- Smith, E. S. (1993). On the shoulders of giants: From Boole to Shannon to Taube; The origins and development of computerized information from the mid-19th century to the present. Information Technology and Libraries, 12(2), 217.
- Vidmar, D. J. (1999). Darwin on the Web: The Evolution of Search Tools. Computers in libraries, 19(5), 22.
- Voorhees, E. M., & Harman, D. (2000). Overview of the sixth text retrieval conference (TREC-6). InformationProcessing&Management, 36(1),3-35.
- White, R. W., & Marchionini, G. (2006). Examining the effectiveness of real-time query expansion. Information Processing & Management, 43 (3), 685-704.

الفصل الثامن

أساليب الاسترجاع

◄ مقدمة

توجد ثلاثة أساليب أساسية لاسترجاع المعلومات هي: البحث، التصفح، والنموذج الهجين من البحث والتصفح. ويعتمد اختيار الأسلوب الملائم لاسترجاع المعلومات على عدة عوامل، لعل أبرزها وأهمها نوع وطبيعة المعلومات التي يحتاج إليها مستفيد بعينه. ويعالج هذا الفصل الأساليب الثلاثة المستخدمة في استرجاع المعلومات من حيث الملامح والتطبيقات.

قام كول (Koll,2000) بتشريح عملية استرجاع المعلومات، حيث أشار إلى أن عملية استرجاع المعلومات هي عبارة عن البحث عن أبرة في كومة قش، حيث إن الإبرة تمثل الوثيقة أو الوثائق التي يبحث عنها المستفيد، وكومة القش هي مجموعة الوثائق المخزنة بقواعد بيانات نظام استرجاع المعلومات.

وقد وضع كول قائمة بالاحتمالات المختلفة لاسترجاع المعلومات من أي نظام وهي كالتالي:

1. البحث عن وثيقة معينة في نظام محدد مثل البحث عن إبرة معينة في كومة قش واحدة.
2. البحث عن وثيقة محددة في نظام غير معروف أو محدد مثل البحث عن إبرة معينة في كومة غير معروفة من القش.
3. البحث عن وثيقة غير معروفة (محددة) ضمن نظام غير معروف مثل البحث عن إبرة غير معروفة في كومة قش غير معروفة.
4. أي وثيقة في نظام محدد مثل البحث عن أي إبرة في كومة محددة من القش.

5. أفضل وثيقة في نظام محدد ـ أقوى إبرة في كومة قش محددة.
6. معظم الوثائق الجيدة في نظام محدد ـ معظم الإبر القوية في كومة قش محددة.
7. كل الوثائق الصالحة المتاحة في النظام ـ كل الإبر القوية في كومة القش.
8. التأكيد على عدم وجود أي وثيقة بالنظام ـ التأكيد على عدم وجود أي إبرة بكومة القش.
9. أي شيء يشبه الوثيقة بالنظام (وثيقة صالحة جزئياً) ـ أي شيء يشبه الإبرة بكومة القش.
10. التنويه بظهور أي وثيقة جديدة بالنظام ـ التنويه بظهور أي إبرة بكومة القش.
11. أين توجد أنظمة استرجاع المعلومات ـ أين توجد أكوام القش.
12. البحث عن الوثائق أو أي منهما ـ الإبر وأكوام القش أو أي منهما.

وتعد القائمة السابقة مجموعة من الاحتمالات الممكنة غير الحصرية للبحث عن الوثائق في أنظمة استرجاع المعلومات، والذي تم تشبيهه بالبحث عن إبرة في كومة قش. ومن الواضح أن البحث هو الأسلوب الملائم لحالات معينة مثل الحالة رقم (1) وأن التصفح يبدو أنه الأسلوب الملائم لحالات أخرى مثل الحالة رقم (12) وأن بعض الحالات في تلك القائمة تحتاج إلى التصفح والبحث معاً مثل الحالة رقم (5).

8.1 الاسترجاع من خلال البحث

Retrieval by searching

يُعد البحث أحد أهم أساليب استرجاع المعلومات والتي يتم معالجتها في الدراسات المختلفة لاسترجاع المعلومات باستخدام مصطلحات متنوعة مثل:

- البحث بقواعد البيانات Databases Searching
- البحث على الخط المباشر Online Searching
- البحث في الفهارس المتاحة على الخط المباشر OPAC Searching

وغيرها من المصطلحات التي تم استخدامها للإشارة إلى نفس المفهوم، حيث إنه بمجرد أن تتم عملية تمثيل الاستفسار يصبح المستفيد جاهزاً لإجراء البحث لأغراض استرجاع المعلومات من النظام.

8.1.1 ملامح البحث

Characteristics of searching

تسعى عملية البحث عن المعلومات نحو الوصول إلى الوثائق التي تضاهي المصطلحات الواردة باستفسار المستفيد، وذلك من خلال استخدام تقنيات الاسترجاع المختلفة التي تم شرحها في الفصل الخامس. ومن الممكن أن تتم عمليات البحث باستخدام نقاط إتاحة موضوعية Subject Access Point أو نقاط إتاحة غير موضوعية Non subject Access Points. وتشتمل نقاط الإتاحة الموضوعية على الواصفات Descriptors التي يتم اشتقاقها من المكانز، أرقام التصنيف التي يتم استخراجها من خطط التصنيف، رؤوس الموضوعات التي تشتق من قوائم رؤوس الموضوعات وغيرها من المحددات الموضوعية الحرة مثل الكلمات المفتاحية، والعناوين والمستخلصات، أو النصوص نفسها بقاعدة بيانات النصوص الكاملة، وتشتمل المصطلحات غير الموضوعية على لغة الوثيقة، سنة النشر، نوع الوثيقة، أرقام تحديد الهوية مثل[1] (ISSN, ISSN, DOI).. إلخ.

وتُعد عملية البحث نموذجاً فعالاً لاسترجاع المعلومات في حالة الاستفسارات المحددة التي يدرك فيها المستفيد الحاجة إلى الوصول إلى كل الوثائق التي نشرها نجيب محفوظ مثلاً خلال عقد السبعينات، فإن عملية البحث باسم المؤلف، تاريخ النشر سوف تؤدي استرجاع النتائج المتوقعة من النظام. أما إذا كان المستفيد بحاجة إلى معرفة كل من أسهم في تطوير مجال استرجاع المعلومات، فإن البحث وحده قد لا يكون وسيلة ملائمة لتلبية احتياجاته ولا بد أن يقوم أيضاً بالتصفح.

---

(1) ISBN – Intenrational Standard Book Number
ISSN – International Standard Serial Number
DOI – Digital Object Identifier

يعتمد أسلوب البحث عن المعلومات على استخدام تقنيات البحث، مثلاً الاعتماد على المنطق البوليني Boolean Logic والذي يتيح للمستفيد إمكانية دمج أكثر من وجه واحد لعملية البحث باستفسار المستفيد عند الحاجة لذلك. وباستثناء أنظمة استرجاع المعلومات على الإنترنت، فإن معظم نظم استرجاع المعلومات تسمح للمستفيد بإجراء تعديلات على الاستفسار من خلال تحديد البحث في نطاق معين مثل حقول البحث. وتُعد عملية البحث عن المعلومات عملية مستقيمة وممنهجة ولا تضع أي أعباء معرفية Cognitive Load على المستفيد.

ويُنظر إلى عملية البحث عن المعلومات على أنها نشاط له بناء محدد وتقل فيه فرص المفاجأة Serendipity (بمعنى اكتشاف شيء مفيد، ولكنه غير متوقع أثناء عملية البحث)، حيث إن النظام يعرض فقط الوثائق التي تضاهي استفسار المستفيد، إضافة إلى أن المستفيد بحاجة إلى التدريب لتعلم مهارات البحث، حيث إنها مهارات مكتسبة تحتاج إلى تعلم وتطور مع الممارسة في نفس الوقت؛ لكي يتمكن المستفيد منها. ويمكن القول إن التعلم والممارسة عمليات مكلفة للغاية مع النظم مدفوعة الكُلفة في مقابل انخفاض الكُلفة في النظم المجانية.

8.1.2 أنواع البحث
Types of searching

يمكن تصنيف عملية البحث إلى عدة أنواع وفقاً لأهداف البحث كالتالي:

النوع الأول: البحث عن وثيقة محددة: فعندما يكون المستفيد بحاجة إلى وثيقة معينة فإن عملية البحث يُطلق عليها البحث عن مادة محددة Known Item search . وتتم عملية البحث عن وثيقة محددة باستخدام محددات بحث مثل المؤلف والعنوان وغيرها من الحقول البحثية. ويُعد البحث عن وثيقة محددة أبرز مثال لنموذج كول (Koll, 2000) البحث عن إبرة معينة في كومة قش محددة.

النوع الثاني: البحث عن موضوع معين Topic search والذي يحتاج إليه المستفيد لأداء بحث في الإنتاج الفكري المتخصص بغرض تحديد ما إذا كان

هناك باحثون آخرون قاموا بإجراء دراسات في هذا الموضوع، أو التعرف إلى كل الدراسات في موضوع معين. وفي الغالب لا يتفاجأ المستفيدون إذا لم يجدوا بحثاً ذا علاقة بالموضوع الذي يبحثون فيه، كما أنهم عادة ما يكونون سعداء بمعرفة أنه لا توجد أي دراسة نشرت في هذا الموضوع حتى الآن، لأن ذلك يعد مؤشراً قوياً على أصالة أبحاثهم.

وقد أطلق العديد من الباحثين على هذا النوع مصطلح البحث السلبي Negative Search مثل (Stielow & Tibbo,1988)، أو كما أطلق عليه كول (Koll, 2000) التأكيد بعدم وجود أي وثيقة في الموضوع أو أي إبرة في كومة القش. وتجدر الإشارة إلى أن هذا النمط من البحث هو النمط الذي تستخدمه مكاتب براءات الاختراع عند فحص أي براءة جديدة للتأكد من أنه لا توجد أي براءة تم منحها في العالم في نفس الموضوع. ويجب أن يكون البحث السلبي عميقاً وشاملاً، بحيث يتأكد المستفيد أنه لا توجد أي وثيقة تعالج نفس الموضوع الذي يسعى إلى البحث فيه.

النوع الثالث: هو نمط البحث بأغراض الإحاطة الجارية والبث الانتقائي للمعلومات Selective Dissemination of Information وقد تم توضيحه بالفصل الأول، حيث أوضح لوهان (Luhn,1961) آليات خدمات الإحاطة الجارية والبث الانتقائي للمعلومات، والتي انتشرت بصورة كبيرة في مجالات التجارة وإداراة الإعمال والمجتمعات العلمية. ويتم في هذه النظم وضع استفسار جاهز ثابت بالنظام، ثم يتم إجراء البحث بطريقة دورية، وعادة ما تتم تلك العملية بطريقة آلية، وقد مثلها كول في القائمة بالبحث عن أي وثيقة جديدة تضاف إلى كوم القش.

النوع الرابع: المزج بين البحث الموضوعي ونقاط الإتاحة غير الموضوعية: حيث يستخدم البحث الموضوعي أو المفاهيمي عندما يكون لدى المستفيد احتياج معلوماتي ويسعى إلى الوصول إلى مجموعة من الوثائق الصالحة في الموضوع. ولكي يتم إعداد الاستفسار يحتاج المستفيد إلى استخدام نقاط الإتاحة الموضوعية التي تم توضيحها في النقطة سابقاً واستكمالها بنقاط الإتاحة غير الموضوعية مثل تحديد نطاق البحث في لغة معينة، تاريخ نشر.. إلخ.

ويوجد العديد من الآليات وتقنيات البحث التي تستخدم لتحديد مدى شمول أو دقة البحث، والتي يتم قياسها بمعدلات الاستدعاء والتحقيق. وقد أوضح كول (Koll,2000) أنه توجد مجموعة من نماذج البحث من وجهة نظر الاستدعاء والتحقيق في القائمة التي حددها وتشمل:

I. البحث عن أي وثيقة في النظام بمعنى أن الاستدعاء منخفض والتحقيق مرتفع.

II. البحث عن أفضل وثيقة واحدة بالنظام بمعنى ارتفاع معدل التحقيق بالبحث.

III. البحث عن معظم الوثائق الصالحة، ما يشير إلى ارتفاع معدل الاستدعاء.

IV. كل الوثائق الصالحة للموضوع بالنظام تحقيق أعلى قيمة استدعاء perfect Recall.

V. ومن الاحتمالات الأخرى في هذا النطاق الوصول إلى معدل تحقيق منخفض ومعدل استدعاء مرتفع عند قياس معدلات الاستدعاء والتحقيق لأي نظام.

بالمقارنة بغيره من أنواع البحث، فإن البحث الموضوعي أو المفاهيمي يُعد أكثر أنواع البحث تعقيداً، نظراً لأنه يحتاج إلى التقييم من جانب المستفيد باستخدام معايير حكم غير موضوعية أو ثابتة في معظم الأحوال، وذلك للحكم على الصلاحية التي تُعد الأساس لقياس معدلات الاستدعاء والتحقيق.

النوع الخامس: البحث بالفقرات Passage Search تمت الإشارة إليه سابقاً في الفصل الأول، حيث يركز هذا النوع من أنواع البحث على استرجاع فقرات من الوثائق تضاهي استفسار المستفيد، وتقوم بعرض تلك الفقرات. ويعتمد هذا النمط على وظائف الفرز والترتيب Filtering functions بصفة أساسية. وعلى الرغم من ظهور هذا النمط خلال السنوات الأخيرة، إلا أنه أظهر إمكانيات كبيرة في دعم المستفيدين وخاصة في تقليل حجم فيضان النتائج المسترجعة إلى جانب تحسين مستويات الدقة والتحقيق في النتائج المسترجعة.

لقد تم في هذا الجزء شرح وتفصيل 5 أنواع من البحث هي: البحث بمادة معروفة، البحث السلبي، البث الإنتقائي للمعلومات، البحث المركب (نقاط بحث موضوعية وغير موضوعية)، والبحث بالجمل. وتجدر الإشارة إلى أنه توجد تصنيفات أخرى لعمليات البحث مثل ما ورد عن كل من (Baeza-Yates and Ribeiro- Neto 1999) حيث قاما بتصنيف البحث إلى فئتين أساسيتين هما: عشوائي Ad-hoc وتصفية Filtering وأياً كان أسلوب تصنيف عمليات البحث، فإن الهدف النهائي هو أن يستطيع المستفيد أن يحدد الأسلوب الملائم للبحث وبناء استراتيجية بحث سليمة تتوافق وتلبي احتياجاته.

8.1.3 استراتيجيات البحث
Search Strategies

تعرف استراتيجيات البحث بأنها عملية تحويل الاستفسار أو الطلب على المعلومات إلى طريقة لإجراء البحث بنظم استرجاع المعلومات. وقد صنف كل من فينشل وهوجان (Fenichel & Hogan, 1981) في العصر الذهبي للبحث على الخط المباشر، استراتيجيات البحث تحت أربع فئات رئيسة هي كالتالي:

8.1.3.1 استراتيجية أعمدة البناء
Building Block Approach

تبدأ استراتيجية أعمدة البناء بالبحث عن مفهوم واحد Single Concept. ومن نماذج استراتيجية المفهوم الواحد ما تم شرحه في عملية البحث عن الاستفسار الخاص بتصفية النزاعات في الشرق الأوسط في جدول رقم (7.2) وفقاً لما تم شرحه في عملية تحليل المفاهيم. ووفقاً لتلك الاستراتيجية يتم البحث عن كل مفهوم على حدة بصورة مستقلة، وبعد البحث عن المفاهيم المستقلة يتم الدمج بين تلك المفاهيم باستخدام معاملات الربط البوليني.

وتعتمد تلك الاستراتيجية على تحليل عمليات البحث المعقدة إلى عمليات أكثر بساطة، ما يتيح للمستفيدين إمكانية تصحيح أو ضبط استراتيجية البحث في الوقت

المناسب أثناء إجراء عملية البحث. من ثم لا يحتاج المستفيد إلى إعادة إجراء البحث بالكامل بسبب وجود خطأ في حرف أو هجاء كلمة في عبارة البحث. بالتالي فإن نموذج أعمدة البناء يقلل من حجم الضغط الذي يوضع على المستفيد، ويتيح له فرصة أكبر للتركيز على التفاعل مع نظام استرجاع المعلومات. ولهذا السبب فإن هذه الاستراتيجية تعد وسيلة مهمة للمستفيدين بصفة عامة لاكتشاف المفاهيم واكتشاف النتائج المرتبطة بها، كما أنها تعد وسيلة مهمة لتعلم كيفية التعامل مع نظم استرجاع المعلومات خاصة للمستفيدين المبتدئين.

◄ 8.1.3.2 استراتيجية كرة الثلج
SnowBall Strategies

تعرف استراتيجية كرة الثلج أيضاً باستراتيجية استخدام الاستشهادات في حصاد اللؤلؤة (Fenichel & Hogan, 1981) حيث تساعد تلك الاستراتيجية على زيادة أعداد المصادر المسترجعة كما هو الحال في نمو كرات الثلج في وقت نزول الثلج. ومن الواضح أن هذا النموذج يسعى إلى زيادة معدلات الاستدعاء، حيث إنه وفقاً لهذا الأسلوب يقوم المستفيد بإجراء بحث مبدئي وفقاً للنتائج المسترجعة ثم يقوم بتعديل الاستفسار. وتعتمد عملية التعديل على مراجعة وفحص النتائج المسترجعة واختيار المصطلحات الملائمة من تلك النتائج من خلال كلمات العناوين والواصفات والكلمات المفتاحية الواردة في النتائج المسترجعة، ثم إعادة استخدامها وتوظيفها في إعداد استراتيجية أكثر إحكاماً. ومن الممكن أن تتم تلك العملية أكثر مرة، بحيث يتم في كل مرة مراجعة المصطلحات المستخدمة وتعديل الاستراتيجية حتى يصل الباحث إلى أعلى مستويات الرضا عن النتائج المسترجعة.

فعلى سبيل المثال إذا قام مستفيد بالبحث عن موضوع الكتب الإلكترونية Eiectronic Books وقام النظام باسترجاع وثائق عن Stephen King ووثائق تستخدم مصطلح ebooks يقوم المستفيد باستخدام استراتيجية كرة الثلج بتعديل استراتيجية البحث ووضع تلك المصطلحات بالاستراتيجية الجديدة، بغرض توسيع نطاق البحث والحصول على كل النتائج الممكنة في هذا الموضوع. فالمتخصص في

مجال الكتب الإلكترونية يعلم أن Stephen king أول مؤلف شهير يقوم بنشر كتابه في صورة إلكترونية، كما يعلم أيضاً أن مصطلحات ebook, EPUB هي اختصار للمصطلح الكامل electronic book لذلك يجب تضمينها في عملية البحث. وتعتمد استراتيجية كرة الثلج في جوهرها على استخدام إمكانيات البحث المتقدم، وتسعى إلى توسيع نطاق الاستفسار Query Expansion اللذين تمت مناقشتهما سابقاً. وقد أطلق كورفهج (Korfhag, 1997) على هذه الممارسة معالجة الاستشهادات. ويمكن القول إن استراتيجية كرة الثلج تُعد استراتيجية مفيدة في حالة حاجة المستفيد الذي يحتاج إلى دعم لتحديد المصطلحات المرتبطة بالموضوع الذي يبحث عنه بغرض توسيع نطاق البحث.

8.1.3.3 استراتيجة التجزيء المتوالي
The Successive Fraction Approach

تُعد استراتيجية التجزيء المتوالي النموذج العكسي لاستراتيجية كرة الثلج، حيث تبدأ عملية البحث وفقاً لتلك الاستراتيجية باستخدام المفاهيم العريضة Broad Concept ثم يتم تضييق نطاق البحث بطريقة متتالية وفقاً لما سيتم اكتشافه من نتائج، وذلك باستخدام محددات البحث المختلفة مثل معاملات الربط البوليني والتقارب عند صياغة عبارة البحث. فكما سبقت الإشارة إلى معاملات الربط البوليني فإن المعامل NOT يستخدم لاستبعاد مصطلحات من عبارة البحث، كما يستخدم المعامل AND في تحديد نطاق البحث بالربط بين منطقة التماس أو التداخل بين المفاهيم. ويستخدم المعامل with أيضاً في تضييق نطاق البحث من خلال تحديد موضع المصطلحات في العبارة البحثية، والتي يجب أن ترد معاً. ومن أساليب التحديد أو تضييق نطاق البحث استخدام المحددات غير الموضوعية Non Subject Attributes مثل لغة أو نوع أو سنة نشر الوثيقة. ويُعد التحديد باستخدام المحددات غير الموضوعية أكثر سهولة من استخدام المعاملات التي تربط بين المفاهيم في التحديد. نفترض أن باحثاً يريد البحث عن موضوع تصفية الويب Web Filtering كموضوع جدلي Controvensy وليس كموضوع تكنولوجي وبدأ البحث بالمصطلح

تصفية Filtering بالطبع فإن النتائج سوف تتضمن كل شيء له علاقة بالتصفية أو الفلترة يشتمل عليه نظام استرجاع المعلومات. في هذه الحالة لابد من استخدام استراتيجية التجزيء المتوالي لكي يتم الوصول إلى الهدف المحدد من جانب المستفيد. فعلى سبيل المثال في هذه الحالة يتم إضافة المصطلح الويب web لعبارة البحث باستخدام المعامل AND لتصبح عبارة البحث Filtering AND Web: كما يمكن إضافة المصطلح Controversy بنفس الطريقة بعد إجراء البحث بالعبارة السابقة والنظر في حجم النتائج المسترجعة ومدى تطابقها مع احتياجات المستفيدين. على أن يتم استخدام المعامل NOT في تلك الاستراتيجية لتصبح كما يلي: **Filtering AND Web Not Controversy**

بالتالي يتم استبعاد أي نتائج ذات علاقة بمصطلح تكنولوجيا المعلومات information technology. ومن الممكن تحديد عملية البحث بصورة أكثر تفصيلاً للوثائق التي نشرت بين عامي 1990 ـ 2000 من خلال استخدام المعامل AND لتصبح استراتيجية البحث (جدول 8.1) تقسيم المفاهيم والربط بينها وفقاً لاستراتيجية التجزيء المتوالي:

Search Field	Operator	Search Term	المصطلح البحثي
key words		Filtering	الفلترة
key words	AND	Web	الويب
Publishing Year 1990- 2000	NOT	Controversy	الجدل

فكما أوضحنا يسعى نموذج استراتيجية التجزيء المتوالي إلى تضييق نطاق البحث خطوة بخطوة باستخدام إمكانيات التحديد والتضييق المتاحة بنظم استرجاع المعلومات. ويتطلب هذا الأسلوب أن يكون المستفيد على دراية وأن يتم تدريبه وتأهيله على آليات وإمكانيات تضييق نطاق البحث المتاحة بنظم استرجاع المعلومات، إلى جانب تدريبه على التفاعل مع النظام أثناء عملية البحث. وتتطلب

عملية التفاعل مع النظام أن يقوم المستفيد بالاطلاع على عناوين ومستخلصات النتائج المسترجعة في كل دفعة من دفعات البحث لتحديد مدى مطابقاتها للمفهوم الذي يبحث عنه، أم أنه توجد حاجة إلى تضييق أو توسيع المفهوم. وتجدر الإشارة إلى أن الكُلفة كانت عاملاً مؤثراً في استخدام ذلك النوع من عمليات البحث في النظم المتاحة على الخط المباشر Online System خلال السبعينات والثمانينات من القرن الماضي. ونظراً لحاجة المستفيد إلى التعامل مع النظام لفترات طويلة كانت عملية البحث في ذلك الوقت تتم عبر خطوط الهاتف الدولية، ما كان يمثل أكبر عناصر الكُلفة في تلك النظم، إلا أن ظهور الإنترنت وانتشار استخدامه في إتاحة عمليات البحث بقواعد البيانات، قلل من تلك التكاليف بصورة كبيرة، حتى أصبح وقت عملية الاتصال عنصراً غير مؤثر في الكُلفة على الإطلاق. ومازالت كل نظم استرجاع المعلومات تعتمد بصورة كبيرة على إمكانيات تحديد نطاق البحث لتيسير استراتيجية التجزيء المتوالي.

◄ 8.1.3.4 استراتيجية الوجه الأكثر تحديداً
The most Specific Facet Strategy

تستخدم استراتيجية الوجه الأكثر تحديداً كاتجاه أولى مع الاحتياجات البحثية متعددة الأوجه (Fenichel & Hogan, 1981) وتفترض تلك الاستراتيجية أن المستفيد يعرف جيداً كل أوجه الموضوع الذي يبحث عنه ويستطيع تجزئته إلى مجموعة مفاهيم تتضمنها العبارة البحثية، ثم يقوم بتحديد أكثر تلك المفاهيم أهمية وتحديداً. وتعد تلك الاستراتيجية من أكثر الاستراتيجات كفاءة، حيث إنها تستغرق أقل قدر من الوقت، نظراً لأن المستفيد يبدأ عملية البحث بأكثر المفاهيم تحديداً. ويرجع ذلك إلى أن نتائج البحث عن أكثر المصطلحات تحديداً تساعد المستفيد في التعرف إلى الحجم المتوقع للنتائج في تلك الاستراتيجية، فقد يكون من غير المنطقي الاستمرار في البحث بنفس الاستراتيجية إذا كان البحث بأكثر المصطلحات تحديداً يسترجع عدداً محدوداً من النتائج أو لا يسترجع أي نتائج على الإطلاق، حيث إن ذلك سوف يؤدي إلى استراتيجية صفرية (zero strategy) أي تسترجع صفراً من

النتائج) أو استراتيجة الندرة Strategy of Few التي تسترجع عدداً محدوداً من النتائج لكي تفي باحتياجات المستفيد، وذلك في مقابل استراتيجية الوفرة، The strategy of Abundance، فعلى سبيل المثال الموضوع التالي يشتمل على ثلاثة أوجه رئيسة:

Treatment of prognosis of neuroendocrine trumors

من ثم يكون الموضوع جرعات علاج أورام الغدد الصم عصبية بالرئة، وهو كما يتضح موضوع معقد ومن ثم نلاحظ أن هذا الموضوع ينقسم إلى ثلاثة أوجه رئيسة هي كالتالي:

الوجه الأول: جرعات علاج Treatment and prognosis

الوجه الثاني: أورام الغدد الصم عصبية neuroendocrine trumors

الوجه الثالث: الرئة lung

ومن بين هذه الأوجه الثلاثة يتضح أن موضوع أورام الغدد الصم عصبية هو الموضوع الأكثر أهمية والأكثر تحديداً، ويجب أن يتم البحث به أولاً وفقاً لهذه الاستراتيجية. فإذا استرجع البحث بالمصطلح neuroendocrine trumors وثيقتين فقط على سبيل المثال، فإنه من المحتمل ألا يسترجع البحث بعد إضافة الأوجه الأخرى أي وثائق أخرى، ما يؤدي إلى استراتيجية صفرية، حيث إن البحث بالمصطلحات الثلاثة باستخدام معامل الربط AND الملائم لتلك الأوجه لن يسترجع بأي حال من الأحوال أكثر من وثيقتين، إلا أنه من المحتمل أن يسترجع عدداً أقل من الوثائق؛ واحداً أو صفر. وعلى الرغم من كفاءة هذا النوع من أنواع استراتيجيات البحث، إلا أنه نموذج في غاية التعقيد، نظراً لأنه لا يمكن إنكار مدى تعقيد عملية التحليل المفاهيمي التي يتضمنها وخاصة التركيز على أكثر المفاهيم تحديداً، ما يجعله نموذجاً صعباً بالنسبة للمستفيد المبتدئ والبسيط؛ حيث إن عملية تعيين أكثر المفاهيم تحديداً من الاحتياجات متعددة المفاهيم تعد عملية معقدة إلى حد ما. لذلك فإننا لا نوصي باستخدام هذا النموذج من جانب المبتدئين في عمليات البحث واسترجاع المعلومات.

ومن الممكن عمل امتداد لاستراتيجة المفهوم الأكثر تحديداً بالاعتماد على استراتيجية الوجه الثاني الأكثر تحديداً the second most specific face حيث يتم اختيار ثاني أكثر مفهوم تحديداً في حالة تعذر التعامل مع المفهوم الأول ويتم استخدامه في إجراء البحث. ومع ذلك فإن هذا النوع من الاستراتيجيات نادراً ما يتم استخدامه من جانب مجتمع المستفيدين من نظم استرجاع المعلومات. وقد قام كل من فينشل وهوجان (Fenichel & Hogan, 1981) بوصف هذه الاستراتيجية باستخدام مصطلح اتجاه الندرة أولاً the lowest first approach في إشارة إلى أن الوجه الأكثر تحديداً عادة ما يسترجع أقل عدد من النتائج.

8.1.4 نحو الاستراتيجية الأكثر ملاءمة وسرعة

سبقت الإشارة إلى أن كل استراتيجيات البحث تم بناؤها وتطويرها في وقت انتشار ونمو النظم المتاحة على الخط المباشر، وأن هذه الاستراتيجيات لم تعد ملائمة للبيئة الرقمية الجديدة. فقد شهدت بيئة استرجاع المعلومات تغييرات كبيرة مع ظهور نظم استرجاع الإنترنت، كما أن المستفيدين أنفسهم حدثت لهم تغييرات كبيرة، حيث اختفى دور وسيط المعلومات الذي كان يقوم بالبحث نيابة عن المستفيد النهائي، وأصبح المستفيد يتفاعل بصورة مباشرة مع أنظمة استرجاع المعلومات. وعلى الرغم من أن استراتيجيات مثل أعمدة البناء واستراتيجية كرة الثلج لاتزال من أكثر الاستراتيجيات تفضيلاً من جانب قطاع كبير من المستفيدين، إلا أن القطاع الأكبر من المستفيدين يفضل البحث بكلمة واحدة أو مجموعة كلمات دون استخدام أي محددات أو علاقات وروابط بولينية فيما بينها وهو النموذج الذي تعتمد عليه محركات بحث الإنترنت، التي تستخدم نموذج البحث السريع من خلال صندوق بحث بسيط (Jansen, Spink & Saracenvic, 2000; Siegfried, Bates& Wilde, 1993) وفي المقابل نجد أنه نادراً ما يستخدم المستفيدون من نظم استرجاع المعلومات اليوم استراتيجية التجزيء المتوالي أو استراتيجية الوجه الأكثر تحديداً أولاً. ومن ناحية أخرى نجد أن بعض أنظمة استرجاع المعلومات تستخدم بعض الإمكانيات الأساسية في البحث مثل نوع لغة (مضبوطة أم لغة طبيعية) في الأنظمة

التي تستطيع توفير آليات بحث متنوعة، معاملات الربط البوليني وتدمجها في واجهات استرجاع المعلومات الحديثة، حيث يمكن للمستفيد أن يحدد خيارته في البحث من خلال نماذج البحث Search Forms أو الأزرار المجهزة مسبقاً Predefind Buttons أو القوائم المنسدلة Drop Down Menus دون الحاجة إلى كتابة تلك الخيارات في صندوق البحث.

وعلى الرغم من تنوع آليات البحث وتعددها وابتكار العديد من الأساليب التي تمكن المستفيد من الوصول والاكتشاف، إلا أن آليات وإمكانيات البحث المتنوعة تقف قاصرة عن تلبية العديد من الطلبات المعرفية للمستفيدين وفقاً لإمكانياتهم وقدراتهم البحثية والتي يجب أن يراعيها أي نظام استرجاع معلومات، ما اضطر الباحثين في مجالات استرجاع المعلومات إلى البحث عن آليات توفر بدائل للمستفيدين في الوصول إلى مصادر المعلومات. وتمثلت تلك الآليات في الاسترجاع بالتصفح والذي سيتم عرضه في الجزء التالي.

8.2 الاسترجاع بالتصفح

Retrieval By Browsing

يعد التصفح أحد أهم أساليب استرجاع المعلومات، على الرغم من أنه لم يلقَ الاهتمام الكافي من جانب المهتمين باسترجاع المعلومات، مقارنة بالبحث حتى الثمانينات والتسعينات من القرن الماضي، والتي شهدت نمو وانتشار أنظمة الأقراص المدمجة، والفهارس المتاحة على الخط المباشر، إلى جانب بيئة الروابط الفائقة في الشبكة العنكبوتية العالمية. وقد أدى انتشار تلك التقنيات إلى اكتساب التصفح شهرة واسعة وبسرعة كبيرة، حيث أصبح يمثل جدوى اقتصادية في عمليات استرجاع المعلومات. في الوقت الذي تغيرت فيه أساليب الاتصال من النظم المتاحة على الخط المباشر التي كانت عملية الاتصال التليفوني فيها مكلفة جداً إلى نظم استرجاع المعلومات من خلال قواعد البيانات المتاحة على الإنترنت، من ثم أصبحت كل نظم استرجاع المعلومات تتيح التصفح كأحد وسائل الوصول التي تيسر للمستفيدين القيام بهذا الدور.

8.2.1 ما هو التصفح

التصفح هو عملية اختيار المعلومات الملائمة لاحتياجات المستفيدين من خلال قوائم عامة باستخدام آليات القراءة بالقشط والمسح وغيرها من الأنشطة المشابهة. ويسعى المستفيدون إلى استخدام التصفح وسيلةً لاسترجاع المعلومات للحصول على ما يلي:

1. معلومات عن موضوع غير محدد بدقة فيتم الإحالة إلى تعريفه.
2. معلومات عن موضوع من الصعب تخصيصه أو معرفة مجاله بوضوح، مثال ما هي الفئة التي ينتمي إليها هذا الموضوع؟ ومن الممكن في هذا الإطار أن يتم تطوير آلية دولية لبناء شبكة اجتماعية لتعريف الكيانات ووضعها في فئات تحدد مجالها ومداها المعرفي.
3. معلومات عامة عن الموضوع و/ أو الموضوعات التي يغطيها نظام استرجاع المعلومات.
4. مساعدة المستفيد على الاختيار من بين مزيج من المواد الصالحة وغير الصالحة.
5. اكتشاف والتعرف إلى المواد الجديدة التي يتم إضافتها إلى قواعد البيانات.

وقد تناول مارشونيني ووايت (Marchionini & White, 2007) موضوع التصفح بصورة أكثر تفصيلاً، واستعرض أهميته والحاجة إليه، إلى جانب شرحه بصورة أكثر عمقاً، وأشار إلى أن كل المتطلبات السابقة تمثل أهمية كبرى للمستفيدين، ولكي تستطيع نظم استرجاع المعلومات أن تخدم المستفيدين بكفاءة فإن عليها أن توفر إمكانيات التصفح التي تساعد المستفيدين على الوصول إلى كل ما سبق. وقد أوضح كول (Koll, 2000) أنه في حالات التعامل مع الإبرة في كومة قش أو الأكوام نفسها Needles or hystacks أو ما شابه، فإن التصفح يُعد الوسيلة الأفضل لاسترجاع المعلومات.

فعند التصفح لا يحتاج المستفيد إلى التعبير عن المشكلة المعلوماتية في صورة اصطلاحية محددة باستخدام عبارة بحثية، كما هو الحال في عملية البحث. فعملية

التصفح تحتاج إلى جهد معرفي أقل بكثير مما تحتاج إليه عملية البحث. وذلك رغم أنه أثناء عملية البحث يجب أن يظل المستفيدون على اتصال وتفاعل دائمين مع نظم استرجاع المعلومات بغرض فحص وقياس أو تقييم المعلومات من خلال عمليات التصفح بالقشط أو المسح ثم إصدار أحكام صلاحية عن مدى دقة المعلومات المسترجعة، ما يجعل الحمل المعرفي Congnitive load الذي يبذله المستفيد في عملية تصفح النتائج المسترجعة أكبر بكثير من اختيار البحث كوسيلة لاسترجاع المعلومات. وعلى عكس البحث، فإن التصفح عملية حدسية لا تحتاج إلى تدريب أو خبرة كوسيلة لاسترجاع المعلومات. وقد أوضح مارشيونيني ووايت (Marchionini & White, 2007) أن عملية التصفح هي عملية طبيعية، نظراً لأنها توافق نظرة الإنسان للمصادر الطبيعية والعاطفية والمعرفية، وتتسق مع رؤيته ومراقبته للعالم المادي والبحث عن العناصر المادية. من ثم فإن عملية التصفح تتسم بالسهولة كعملية التنفس عند الإنسان. وعلى الرغم من ذلك فإنه توجد بعض الآليات المتطورة التي تيسر عملية التصفح، وعادة ما يعاني المستفيد من مشكلة عدم وجود إرشادات كافية تمكنه من الاعتماد عليها لمعرفة متى يحتاج إلى الاستمرار في عملية التصفح ومتى يجب أن يتوقف عن تصفح مصدر معين؟ وهذا أمر يشبه القرار الذي يتخذه المؤلف عند التحول من القراءة إلى الكتابة، حيث إن عمليات البحث والتصفح والحاجة إلى الاستمرار في القراءة والتوقف لبدء الكتابة أو الاستمرار في الكتابة والتوقف والتوجه نحو النشر كلها عمليات معرفية تحتاج إلى قرارات شخصية وتعد مؤشراً قوياً للنضج المعرفي لدى الشخص، كما أنها أمور ترتبط بالإشباع المعرفي Knowledge Satisfaction.

وتجدر الإشارة إلى أن التصفح قد يكون نشاطاً فعالاً في كثير من الأحيان، حيث يقود المستفيد بالمصادفة للوصول إلى معلومات لم تكن متوقعة، فالتصفح يتيح للمستفيد البحث عن المعلومات بصورة عشوائية وبطريقة غير مهيكلة في ذهنه مقدماً، ولا توجد عبارة بحثية محددة، وكأي نشاط من أنشطة التفاعل مع المعلومات، فإن التصفح له العديد من المزايا والكثير من العيوب، كما أن له آليات متنوعة ومتعددة سيتم عرضها فيما يلي:

8.2.2 أنواع التصفح

فكما أشرنا من قبل فإن عملية التصفح تُعد طريقة غير مهيكلة لاسترجاع المعلومات، ويقصد بعدم الهيكلة أن المستفيد ليس لديه تصور واضح لهيكل المعلومات، كما يفتقر إلى التحديد الاصطلاحي والعبارة البحثية الواضحة التي يتم صياغتها في صورة استراتيجية بحث. وقد صنفت العديد من الدراسات أنواع التصفح ومنها (eg: Herner,1960; Kowalski,2007; Marchionini & White, 2007). ولعل أبرز هذه التصنيفات تصنيف هرنر (Herner,1960)، الذي صنفها إلى ثلاث فئات هي:

- **التصفح المباشر Direct Browsing** ويقصد به التصفح من أجل الوصول إلى مادة أو مواد محدد ومعروفة.
- **التصفح شبه المباشر Semi Direct Browsing**: يقصد به التصفح من أجل الوصول إلى مادة أو مواد قريبة من صور ذهنية شبيهة من مادة معينة في ذهن المستفيد.
- **التصفح غير المباشر Non Direct Browsing** وقد أشار إليه هرنر بالتصفح العشوائي الذي يقوم به المستفيد من أجل الوصول إلى أي معلومات ذات علاقة بموضوع معين دون أن يكون لدى المستفيد صورة ذهنية محددة أو شبه محددة لما يحتاج إليه أو ما يمكن أن يصل إليه.

كما قام مارشيونيني ووايت (Marchionini & White, 2007) بتصنيف التصفح إلى ثلاث فئات شبيهة لتصنيف هرنر هي:

- النظامي Systematic
- الاستكشافي Exploratory
- العرضي أو غير النظامي Casual or Non -systematic

يستخدم التصفح المباشر أو النظامي عندما يكون المستفيد على علم تام بما يبحث عنه مثل التصفح من أجل الوصول إلى صفحة معينة بأحد المواقع أو

الوصول إلى كلمة محددة في قاموس، بينما يستخدم المستفيدون التصفح شبه المباشر أو الاستكشافي عندما لا يكون لديهم أهداف دقيقة واحتياجات محددة. وتظهر هذا الحالة في مرحلة استكشاف جوانب الموضوع من خلال البحث، فيقوم المستفيدون بعمليات القشط والمسح لتحديد ما يبحثون عنه. فعلى سبيل المثال، قد يكون المستفيد على علم بأن أحد التقارير قد ناقش موضوعاً أو فكرة مهمة، فيقوم المستفيد بتصفح التقرير للوصول إلى تلك الفكرة وتحديدها وفقاً لما ورد بالتقرير، دون أن يكون على علم مسبق بها. ويعد التصفح غير المباشر أو العرضي أقل أسلوب من أساليب التصفح تماسكاً، حيث لا يمكن التنبؤ فيه بما سيصل إليه المستفيد أو مكان وجوده. ويتسم هذا الأسلوب بأنه ليس له احتياجات معلوماتية محددة، وأبرز نموذج لذلك عندما يقوم المستفيد بالقفز من خبر إلى آخر عند مسح موقع للأخبار أملاً في الوصول إلى شيء مفيد يمكن أن يقرأه. وهو مثل ما يحدث مع الباحثين عند تصفح مجلة بموضوعات تدخل في نطاق اهتمامهم.

وقد حدد كوالسكي (Kowalski, 2007) ثلاثة أساليب يقوم بها المستفيدون لتصفح النتائج التي يحصلون عليها:

- **التصفح وفقاً للترتيب**

Browseng By Ranking

تستعرض معظم نظم استرجاع المعلومات في البيئة الرقمية النتائج في صورة مرتبة بالاعتماد على خوارزمية ترتيب محددة، ويسعى كثير من المستفيدين إلى تصفح النتائج ذات علاقة الصلاحية الأقوى بموضوعاتهم أولاً، من ثم يقومون باختيار النتائج مرتبة وفقاً للصلاحية.

- **التصفح بالمنطقة**

Browsing By Zone

عادة ما يتم وضع المعلومات التي لها أهمية خاصة لدى المستفيد في مناطق محددة عند عرض النتائج مثل حقول البيانات التقليدية (العنوان، المستخلص، تاريخ النشر.. الخ) حيث تشتمل هذه الحقول على مواضع معلومات غنية يسعى المستفيد إلى تصفحها.

- **التصفح بالمناطق البارزة**

Browsing By Highlighted Zone

تقوم بعض نظم استرجاع المعلومات بتسليط الضوء على معلومات معينة مثل المصطلحات وعبارات البحث والسياقات التي ترد فيها لكي تساعد المستفيدين على تحديد وإيجاد ما يبحثون عنه بسرعة وفعالية أكبر، لذلك تُعد هذه المناطق البارزة من المناطق المهمة للتصفح.

إضافة إلى ما سبق فقد أشار كوالسكي (Kowalski, 2007) إلى فئتين أساسيتين للتصفح تُستخدمان بكثافة في بيئة استرجاع المعلومات على الإنترنت وهما:ـ

- التصفح بالفئات Browsing By Category
- التصفح بالروابط الفائقة Browsing By Hyper links

وقد برز التصفح بالفئات في أدلة بحث الويب Web Directories مثل ياهو، ففي هذه النوعية من أدوات بحث الإنترنت يتم تجميع المعلومات وتصنيفها تحت فئات محددة مسبقاً بناء على آليات التكشيف والتصنيف للفئات مثل الحاسبات، التعليم، الترفيه، الرياضة. فعلى سبيل المثال المستفيد الذي يبحث عن فيلم لكي يشاهده سوف يقوم طبيعياً بتصفح فئة الترفيه. ويعد التصفح بالروابط الفائقة أحد السمات المهمة التي تتميز بها بيئة الويب، والذي يعد الملمح الأساسي في كل الأنشطة والخدمات المتاحة من خلال بيئة الشبكة العنكبوتية.

وتُعد الروابط الفائقة وحدات طرفية Nodes ومؤشرات Pointers يتم وضعها ضمن النصوص الفائقة بحيث تحاكي بصورة ذكية طريقة التفكير العلائقي Associative Thinking لدى الإنسان، حيث أوضح بوش (Bush,1945) أن عقل الإنسان يعمل بطريقة علائقية، فمع استيعاب نقطة ما تنفجر Snap معها في نفس الوقت إلى نقطة أخرى تقترحها من خلال ترابط الأفكار Association of Thoughts في تطابق مع بعض العقد العنكبوتية المتشابكة لمحاولات خلايا المخ فك ذلك التعقيد.

وقد تحول هذا النمط من التفكير الإنساني إلى واقع ملموس بشكل واضح وعميق

مع اختراع وتطبيق الروابط الفائقة من خلال تيم بيرنرلي. فالويب بأكملها تتكون من معلومات نصية ووسائط متعددة يتم ربطها معاً في روابط فائقة. وتساعد تلك الروابط الفائقة على توجيه المستفيد لتصفح وإيجاد المعلومات الرقمية المتاحة على الويب. لذلك فإن نظم استرجاع المعلومات ذات البنية الفائقة Hyper Structuned IR System أصبحت إحدى أبرز إن لم تكن أهم بيئات تصفح المعلومات الحالية.

ويتضح مما سبق أن التصفح يمكن تصنيفه إلى عدة فئات باستخدام معايير متعددة، إلا أن الهدف من التصفح لابد أن يظل واحداً في نظام استرجاع المعلومات وهو تيسير الوصول إلى المعلومات التي يسعى إليها المستفيد.

8.2.3 استراتيجيات التصفح
Browsing Stratagies

يُعد التصفح أحد آليات الوصول إلى المعلومات، مثله في ذلك مثل البحث، وتتم عمليات التصفح من خلال استراتيجيات متنوعة. وقد أوضح مارشونيني ووايت (Marchionini & White 2007) أنه توجد أربع استراتيجيات للتصفح هي: المسح، الملاحظة، الإبحار، المراقبة.

8.2.3.1 المسح Scan

يُعد المسح أكثر استراتيجيات التصفح تنظيماً نظراً لأنه يتعامل مع الكيانات المحددة تحديداً دقيقاً في بيئات استرجاع المعلومات عالية التنظيم. فالمستفيد الذي يستخدم تلك الاستراتيجية يعرف بالضبط ما الذي يبحث عنه، حيث يبحث عن كيان محدد الهوية، من ثم فإنه يمسح المعلومات التي يتيحها النظام إما خطياً Linearly أو اختيارياً Selectively. ويتم المسح الخطي من خلال تصفح فضاء المعلومات باستخدام آلية التتابع الخطي Sequential Linearly التي يقوم فيها المستفيد باستعراض المواد مادة مادة (Marchionini & White, 2007)ومن أبرز الأمثلة على ذلك مسح قائمة عناوين النتائج المسترجعة للوصول إلى المادة المطلوبة.

أما المسح الاختياري فيعني استعراض أجزاء محددة من المعلومات (على سبيل المثال الرؤوس، الروابط، الصور، والمحتوى المتاح بألوان مختلفة بمواقع الويب) دون غيرها من المعلومات التي يعرضها النظام. ويقوم المستفيد بمسح هذه العناصر لتحديد الفئات التي يرغب في الحصول عليها والاختيار من بينها. فقد يبحث المستفيد عن موضوع معين ويحتاج فيه إلى استعراض الصور أو الوسائط المتعددة. وتُعد استراتيجية المسح الاستريجية الأساسية التي تستند إليها آليات التصفح النظامي Systematic Browsing والتي يتم تطبيقها في أدوات تصفح الإنترنت.

◄ 8.2.3.2 الملاحظة Observation

مقارنة بالمسح تُعد «الملاحظة» استراتيجية التصفح الرئيسة التي تستخدم في عمليات الاكتشاف أو التصفح العام Casual Browsing حيث يجب أن يكون المستفيد منتبهاً إلى الأجزاء التي يتم عرضها ويكون على وعي بالمعلومات وبالأجزاء الأخرى التي يعرضها الموقع مثل الإعلانات حتى لا يتشتت انتباهه. بمعنى آخر أن النظام يعرض العديد من المعلومات للمستفيد، لذلك لابد أن يكون المستفيد منتبهاً ويركز على احتياجاته ويتجاهل المعلومات الأخرى التي لا تدخل في نطاق اهتمامه حتى لا يتشتت في مواقع ليس لها علاقه باحتياجاته المعلوماتية.

◄ 8.2.3.3 الإبحار Navigation

هو من استراتيجيات التصفح التي تسعى إلى تحقيق التوازن بين تأثير المستفيد وبيئة نظام استرجاع المعلومات، حيث تقوم بيئة استرجاع المعلومات بتقييد عملية التصفح في مجموعة من المسارات المحتملة للتصفح ويقوم المستفيد بممارسة التصفح بنفسه من خلال اختيار المسار الذي يتبعه.

كما تعتمد أيضاً استراتيجية الإبحار Navigation على التغذية المرتدة من نظام استرجاع المعلومات، والتي يمكن أن تستخدم بصورة نظامية Systematic أو عرضية Casual أثناء عملية التصفح. وتجدر الإشارة إلى أن استراتيجية الملاحظة غالباً ما يتم تطبيقها مقترنة باستراتيجية الإبحار (Marchionini & White, 2007).

8.2.3.4 المراقبة /المتابعة

هي استراتيجية تشبه استرتيجية المسح، لكنها تتم في البيئات ذات البنية المعلوماتية الفقيرة هيكلياً (Poorly Structured Marchionini & White, 2007). فأثناء قيام المستفيد بتصفح النتائج المسترجعة من النظام، من الممكن أن يقوم أيضاً بمتابعة بعض التقارير الإخبارية التي يبثها النظام من خلال الراديو. وتعتمد تلك الاستراتيجية على فلسفة استخدام المسارات الموازية في البحث عن المعلومات (مسار تصفح المعلومات التي يحتاج إليها المستفيد يتم بالتوازي مع متابعة التقارير الإخبارية التي يبثها الراديو أو التلفزيون). وعادة ما تستخدم استراتيجية المراقبة (المتابعة) في عمليات التصفح الاكتشافي Exploratory Browsing الذي يسعى إلى الوصول إلى تفسيرات وشروح للمفاهيم والكيانات المعرفية.

ويمكن القول بإيجاز إن التصفح يُعد إحدى آليات الوصول إلى المعلومات من خلال الاستعراض والاكتشاف. وتختلف آلية التصفح عن آلية البحث التي تمت مناقشتها سابقاً في مدى تحكم المستفيد في المدخلات وما ينتج عنها. ولا توجد معايير واضحة يمكن للمستفيد من خلالها أن يحدد متى يمكن أن يستمر أو أن يتوقف عن التصفح. ولا توجد مؤشرات يمكن الاستناد إليها من جانب المستفيد في متابعة العمل باستراتيجية معينة أو تغييرها سوى طبيعة بيئة نظام استرجاع المعلومات. وعادة ما يعتمد المستفيدون على عدد من المعايير الكيفية في تحديد الاستراتيجية التي يتبعونها مثل مدى رضا المستفيد والجهد المعرفي المطلوب، وذلك من أجل اتخاذ القرار الملائم لاختيار استراتيجية التصفح الملائمة. كما أنه لا توجد خطوط فاصلة تحدد متى يمكن اختيار أي استراتيجية يتبعها المستفيدون وتحت أي ظرف. هل التصفح كوسيلة استرجاع يعمل بكفاءة أعلى عندما يتم استخدامة مع البحث أم أثناء البحث أم بعد البحث أم ما قبل البحث، أم هل يعمل بكفاءة أعلى إذا تم استخدامه بشكل مستقل؟ ويحاول الجزء التالي الإجابة على هذه التساؤلات.

8.2.4 التكامل بين البحث والتصفح في الاسترجاع

يعد البحث والتصفح أبرز الأساليب الفريدة والمميزة لاسترجاع المعلومات، فمنذ أكثر من نصف قرن مضى قام لوهان (Luhn,1958) بتصنيف طرق الاسترجاع إلى ثلاثة طرق أساسية هي:

1. استرجاع المعلومات من خلال البحث في مصفوفة مرتبة Ordered Array من التسجيلات المخزنة.
2. استرجاع المعلومات من خلال البحث بمصفوفة غير مرتبة Nonordered Array من التسجيلات المخزنة.
3. مزيج من الطريقتين السابقتين.

ومما لا شك فيه أن المصطلحات وبنية نظم قواعد البيانات قد تغيرت كثيراً مع التطورات التي حدثت خلال تلك الفترة. ومع ذلك يمكن النظر إلى الطريقة الأولى التي وصفها لوهان على أنها التصفح، والطريقة الثانية على أنها البحث. أما الثالثة فهي الطريقة التي يجب شرحها بتفصيل حيث تعمل على المزج بين الأسلوبين.

8.2.5 المقارنة بين التصفح والبحث

في الجزء السابق تم شرح ملامح عمليتي البحث والتصفح. وقد أشار كوكس (Cox,1992) إلى أنه يمكن النظر إلى التصفح على أنه يحدد مسار أين إلى ماذا Where To What. وتستند الفكرة الأساسية إلى أن المستفيد يعرف أين يبدأ بقاعدة البيانات ويريد أن يعرف ما المتاح من مصادر بها. وعلى العكس فإن البحث ينطلق من ماذا إلى أين From What to Where وتستند الفكرة إلى أن المستفيد يعرف ما الذي يحتاج الوصول إليه وأين توجد تلك المعلومات بقاعدة البيانات. وقد وصف مارشونيني ووايت (Marchionini & White, 2007) البحث بأنه الاستراتيجية الرسمية والتحليلية للوصول إلى المعلومات، بينما وصف التصفح بأنه استراتيجية غير رسمية واعتباطية Informal and Heuristic. وبعيداً عن هذا الوصف فإن البحث والتصفح يختلفان عن بعضهما بعضاً في الجوانب التالية:

I. حاجة المعلومات أو الاحتياج المعلوماتي Information Need

تُعد الحاجة إلى المعلومات إحدى أهم المعايير الأساسية التي يمكن على أساسها تحديد الطريقة الملائمة للوصول إلى المعلومات، سواء كانت من خلال البحث أو التصفح. ففي حالة الاحتياجات المعلوماتية المعروفة والمحددة، فإن البحث يظهر كأفضل اختيار للمستفيد، حيث إنه يساعد المستفيد في الوصول إلى ما يحتاج إليه بفاعلية وكفاءة، نظراً لأنه يبحث عن إبرة في كومة القش A Needle from Haystack.

II. وفي المقابل فإن التصفح يُعد البديل الأمثل للمستفيد في حالة الاحتياجات المعلوماتية الفضفاضة (الواسعة) وغير المحددة. ويمكن للمستفيد في هذه الحالة استخدام تكتيكات (آليات) تصفح مختلفة مثل المسح والإبحار لتحديد ما إذا كانت توجد أي معلومات صالحة حول الموضوع الذي يبحث عنه بنظام استرجاع المعلومات أم لا تضاهي احتياجاته. كما أن التصفح يساعد في هذه الحالة على تمكين المستفيد من الوصول إلى التحديد الدقيق لاحتياجاته المعلوماتية والمصطلحات الملائمة لها، ما يساعد على إجراء بحث أكثر دقة وكفاءة.

III. كفاءة وإمكانات التحسين Efficiency and potential for Improrement

عند المقارنة بين البحث والتصفح فإنه يجب أن يؤخذ في الاعتبار كفاءة الاسترجاع وإمكانية تحسين تلك الكفاءة. فيمكن القول بصفة عامة إن البحث سريع Quick ومُركز Focus وموجه مباشرة إلى النقطة Right to The Point التي يحتاج إليها المستفيد، في حين أن التصفح يستهلك وقتاً طويلاً، وغير مركز بدقة على نطاق محدد، كما أنه من المحتمل أن يؤدي إلى تشتت Distracted المستفيد. وعلى الرغم من أنه توجد العديد من الأساليب التي يمكن بها للمستفيد أن يقوم بتضييق نطاق البحث، ما يساعد على تحسين مستوى أداء الاسترجاع، إلا أن ذلك لا يتحقق في التصفح الذي لا يوجد به آليات لتحسين الأداء، إضافة إلى أن المستفيد سوف يحصل على المعلومات فقط من الجزء الذي يتصفحه. ونظرياً يمكن لعملية التصفح أن تستمر إلى ما لانهاية إذا لم يقم المستفيد بوقفها وإنهائها. في نفس الوقت الذي تنخفض فيه دقة عملية التصفح في هذه العملية الممتدة Prolonged Process.

IV. الحمل المعرفي Cognitive Load

يمكن تقسيم عملية البحث عن المعلومات إلى ثلاث خطوات أساسية هي:

- تمثيل الاستفسار Representing the Queny
- إجراء البحث Conducting the Search
- تقييم النتائج Evaluating the Results

وتحتاج الخطوتان الأولى والثالثة حملاً معرفياً كبيراً نسبياً مقارنة بالخطوة الثانية إذا لم يحاول المستفيد أن يتفاعل مع النظام أثناء عملية البحث. وفي المقابل فإن عملية التصفح تمتاز بارتفاع مستوى التفاعل بين المستفيد والنظام. فعملية التصفح سوف تتحول إلى عملية عديمة الجدوى إذا لم يتفاعل المستفيد مع النظام، وظل منتبهاً لما يتم عرضه من النظام. وتجدر الإشارة إلى أن عملية التصفح منهكة للمستفيد الذي يحتاج إلى التركيز لفترات طويلة أثناء عملية التصفح، حيث يحتاج إلى تقييم نتائج التصفح بشكل مستمر وبسرعة وفقاً لمعايير محددة لاختيار البديل المناسب الذي يمكنه من الانتقال إلى المرحلة التالية من التصفح.

لذلك فإن التصفح يعتمد على قدرة المستفيد على تمييز النتائج الصالحة أثناء التصفح مقارنة باستدعاء Recall النتائج الصالحة عند البحث في النظام، ما يضع عبئاً آخر على المستفيد.

V. المصادفة Serendipity

تلعب المصادفة في عملية البحث دوراً محدوداً أو أنها غير موجودة تقريباً، نظراً لأن النظام يضاهي استفسار المستفيد بما هو متاح بقاعدة البيانات. فمن غير العملي أو المحتمل أن يتمكن المستفيد من مسح النظام بأكمله لتمييز المعلومات التي تضاهي استفساره وتحديد ما إذا كانت هناك معلومات إضافية غير التي تم استرجعها من النظام. وفي المقابل فإن التصفح يخضع لاحتمالات المصادفة في الوصول إلى نتائج غير محتملة، حيث إنه من المحتمل أن يصل المستفيد إلى معلومات مفيدة وغير متوقعة عند تصفح النظام.

VI. الجهد :Efforts

تتميز عملية البحث بأنها عملية منظمة لها بنية Structured لذلك يمكن أن يتم تأهيل وتدريب المستفيد عليها بحيث يتمكن من التعامل مع كافة أنظمة البحث، وفي المقابل فإن عملية التصفح هي مجموعة إجراءات طبيعية حدسية تتم من جانب المستفيد ولا تحتاج إلى قضاء وقت في التدريب والتأهيل لتلك العملية، إضافة إلى ذلك فإن عملية التصفح لا تحتاج إلى تمثيل الاستفسار، ما يُحرر المستفيد من مهمة صعبة معقدة جداً تتمثل في تحديد المصطلحات البحثية والربط بينها واختيار آلية البحث المناسبة. ويساعد تحرر المستفيد من كل هذه المهام المعقدة على التركيز أكثر على عملية التصفح.

ويلخص الجدول8.1 عناصر المقارنة بين البحث والتصفح والتي تتضمن خمسة محاور أساسية:

جانب المقارنة العملية	الاحتياج المعلوماتي	الكفاءة	الحمل المعرفي	المصادفة	الجهد المطلوب
البحث	محدد ومعروف	مرتفع	خفيف	أقل	أكبر
التصفح	واسع وغير مؤكد	منخفض	ثقيل	أكبر	أقل

◄ 8.3 النهج المتكامل
Integrated Approach

أوضحت المقارنة الواردة في الجدول 8.1 أن لكل من البحث والتصفح مزايا وعيوباً. فكل منهما يعمل كطريقة استرجاع مثالية في ظروف معينة وبشروط محددة. ذلك على الرغم من أنه توجد بعض المواقف التي يبدو فيها أن هناك نهجاً أو طريقة أكثر ملاءمة من الأخرى، فإن تحقيق التكامل بينهما يؤدي إلى تحسين أداة الاسترجاع بصفه عامة. فمن الممكن ألا نحتاج إلى إجراء بحث في بعض الحالات، إلا أن التصفح يبدو أنه نشاط أساسي في كل عمليات الاسترجاع من أجل الحكم على صلاحية

النتائج المسترجعة. علاوة على ذلك، فإن أنظمة استرجاع المعلومات تم تصميمها لتحفيز وتشجيع المستفيدين على النهج المتكامل في مجتمع استرجاع المعلومات.

ففي أنظمة استرجاع المعلومات التي تم تصميمها مع بدايات ظهور نظم الاسترجاع على الخط المباشر تم استخدام القوائم Menus وخيارات البحث Search Options بشكل متوازٍ، من ثم يمكن للمستفيد أن يختار البحث أو التصفح حسب احتياجاته. ومع بداية نظم استرجاع المعلومات من خلال الإنترنت ظل النموذج الأساسي لتيسير الوصول إلى المعلومات هو استخدام الأدلة Directories وآليات البحث Search Mechanism جنباً إلى جنب. وذلك على الرغم من أن بعض النظم التي تم تطويرها وإتاحتها للمجتمع العام في بدايات استرجاع المعلومات من خلال الإنترنت، استخدمت نموذجاً واحداً للوصول إلى المعلومات مثل استخدام ياهو للتصفح من خلال الأدلة واستخدام محرك البحث Altavista للبحث، وليس كليهما.

ولحسن الحظ فإن العديد من أنظمة استرجاع المعلومات على الإنترنت أدركت سريعاً مزايا دعم كل من آليات التصفح والبحث في نظام استرجاع واحد، ما أدى إلى تغيير تصميمها وبنياتها للأنظمة بشكل سريع. لذلك فإنه من الصعب أن تجد نظام استرجاع على الإنترنت لا يوفر آليات لدعم البحث والتصفح معاً في نظام واحد.

ويتمتع المستفيدون بمزايا المنهج المتكامل ليس فقط لوجود كل الأدوات الملائمة للوصول إلى المعلومات، ولكن أيضاً لأن هذا النهج يمكنهم من الوصول إلى معلومات أكثر من نفس نظام الاسترجاع. فعلى سبيل المثال نجد أن ياهو Yahoo يدعم البحث داخل إمكانيات تصفح الفئات التي يتيحها، من ثم يمكن البحث في فئه واحدة مثل Arts، لذلك فإن البحث داخل فئة تصفحية واحدة يشبه البحث في قاعدة بيانات متخصصة في مجال الفئة التصفحية التي يتم البحث فيها. لذلك يمكن القول إن البحث والتصفح نموذجان متكاملان في هذه البيئة. وبطريقة مشابهة فإن نتائج البحث في نظم استرجاع المعلومات اليوم يتم تجميعها آلياً في فئات لتيسير عمليات التصفح وتوسيع وتضييق نطاق البحث. من ثم فإن تطبيق البحث والتصفح في النظم يعطي قيمة مضافة، حيث إن واحد (البحث) مضاف إلى واحد (التصفح)،

من الممكن أن يكون أكثر من اثنين إذا تم دمجهما بحكمة وكفاءة. وتُعد هذه المعادلة صحيحة في إطار النهج المتكامل الذي تم شرحه.

المصادر

- Bush, V. (1945). As we may think. The atlantic monthly, 176(1), 101-108.
- Cox, K. (1992, November). Information retrieval by browsing. In Proceedings of The 5th International Conference on New Information Technology, Hongkong.
- Fenichel, C. H., & Hogan, T. H. (1981). Online Searching: A Primer, Learned Information Ltd.
- Korfhage, R. R. (1997). Information Retrieval and Storage, New York: John Wiley & Sons, 349 p
- Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. Information processing & management, 36(2), 207-227.
- Kowalski, G. J. (2007). Information retrieval systems: theory and implementation (Vol. 1). Springer.
- Koll, M. (2000). Track 3: information retrieval. Bulletin of the American Society for Information Science and Technology, 26(2), 16-18.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.
- Luhn, H. P. (1961). Selective dissemination of new scientific information with the aid of electronic processing equipment. American Documentation, 12(2), 131-138.
- Marchionini, G., & White, R. (2007). Find what you need, understand what you find.International Journal of Human [# x02013] Computer Interaction,23(3), 205-237.
- Ricardo, B. Y. (1999). Modern information retrieval. Pearson Education India.
- Siegfried, S., Bates, M. J., & Wilde, D. N. (1993). A profile of end user searching behavior by humanities scholars: The Getty Online Searching Project Report No. 2. Journal of the American Society for Information Science, 44(5), 273-291.
- Stielow, F., & Tibbo, H. (1988). The negative search, online reference, and the humanities: A critical essay in library literature. RQ, 358-365.

الفصل التاسع

نماذج استرجاع المعلومات

◄ 9 مقدمة

يعرف النموذج Model بأنه وصف دقيق لنظرية أو نظام يأخذ في الاعتبار كل الخصائص والملامح الخاصة بهذا النظام (Soukhanov,et al ,1984). وقد تم تطوير عدة نماذج لاسترجاع المعلومات خلال النصف الثاني من القرن العشرين. ويستعرض هذا الفصل النماذج المختلفة لاسترجاع المعلومات بغرض وضع أساس الممارسة المهنية القائمة على فهم تلك النماذج المختلفة وطرق عملها.

ويمكن تصنيف نماذج استرجاع المعلومات وفقاً لعدة مستويات. وقد اعتمدت الملامح الأساسية للتصنيف على نظريات ومفاهيم تم اشتقاقها من مجالات أخرى، منها على سبيل المثال المنطق البوليني Boolean Logic الفراغ الاتجاهي Vector Space الاحتمال Probability . وقد وضع المتخصصون في استرجاع المعلومات طرقاً وأساليب متعددة لتصنيف كل نماذج استرجاع المعلومات التي تم تطويرها حتى الآن ومنهم (Baeza –Yates & Ribeiro-Neto,1999; Sparck Jones & Willett, 1997).

وقد قام أنجويرسن وجارفلين (Ingwersen, & Järvelin, 2006) بتوسيع نطاق التصنيف والتقسيم إلى فئات لاسترجاع المعلومات الذي وضعه كل من بيلكن وكرافت (Belkin and Craft, 1987) والذي اشتمل على النموذجين الأساسين للمضاهاة وهما المضاهاة التامة Exact Match والمضاهاة الجزئية Best Match.

ويركز هذا الفصل على النماذج الموجهة لخدمة النظم System Oriented Models مثل المنطق البوليني، الفراغ الاتجاهي، الاحتمالات. أما النماذج الأخرى لاسترجاع المعلومات مثل النماذج المعرفية الموجهة للمستفيدين User Oriented Cognitive Model فلن يتم

معالجتها في هذا الكتاب، حيث إنها تميل إلى مجال سلوك البحث عن المعلومات Information Seeking Behavior.

9.1 المضاهاة: أساس كل نماذج استرجاع المعلومات

تعد المضاهاة هي الأساس الذي تعتمد عليه كل أنظمة استرجاع المعلومات رغم أنها ليست نموذجاً إنما هي المكوّن الأساسي لأي نموذج. وقد سبقت الإشارة إلى أن المضاهاة هي الآلية الأساسية في كل أنشطة استرجاع المعلومات. فالمضاهاة يمكن أن تتم بين المصطلحات أو بين مقاييس تشابه Similarity Measuremeuts مثل المسافة Distance أو تردد المصطلحات Term Frequency. وتتم مضاهاة المصطلحات مباشرة على المصطلحات التي تشتق أو تخصص لوصف الوثائق أو الاستفسارات أو غيرهما من أساليب التمثيل التي يتم على أساسها إجراء مضاهاة لمقياس التشابه Similarity Measuremeut Matching بصورة غير مباشرة على المقاييس التي يتم الحصول عليها من تنفيذ العملية الحسابية. على سبيل المثال المسافة بين الزوايا كما هو الحال في نموذج الفراغ الاتجاهي، وسوف يركز القسمان التاليان على مناقشة هذين النوعين من أنواع المضاهاة.

9.1.1 مضاهاة المصطلحات
Term Matching

سبقت الإشارة إلى أن المصطلحات التي تستخدم في تمثيل المعلومات بنظم استرجاع المعلومات تأخذ أشكالاً متعددة مثل الكلمات المفتاحية (Keywords) الواصفات Descriptors، المؤشرات Identifiers. وتشتمل المصطلحات على أشكال متنوعة مثل الكلمات، العبارات أو غيرها من أشكال التعبير مثل المعادلات.. الخ، إضافة إلى ذلك فإن مضاهاة المصطلحات من الممكن أن تتم في أي شكل من الأشكال الأربعة التالية:

- المضاهاة التامة Exact Match

ـ المضاهاة الجزئية Partial Match.

ـ المضاهاة بالموضع Positional Match.

ـ المضاهاة النطاقية Range Match .

وسوف نتناول فيما يلي كل طريقة من طرق المضاهاة وطريقة عملها.

9.1.2 المضاهاة التامة
Exact Match

تعني أن تمثيل الاستفسار Query Representation يضاهي تماماً تمثيل الوثيقة Document Representation في نظام استرجاع المعلومات.

ولعل أبرز نماذج المضاهاة التامة البحث باستخدام الحروف الحساسة Case Sensitivity والبحث بالجمل والعبارات بنظم استرجاع المعلومات. فعلى سبيل المثال مصطلح تصفية أو فرز الويب Web Filtering يمثل استفسار المستفيد ويظهر بنفس الشكل في الوثيقة وبالنظام الذي يتم البحث فيه. من ثم يحصل المستفيد على نتيجة مطابقة تماماً لاستفساره.

9.1.3 المضاهاة الجزئية
Partial Match

على عكس المضاهاة التامة، فإن المضاهاة الجزئية تتعامل مع جزء فقط من مصطلحات الاستفسار والذي يظهر في النتائج المسترجعة والتي تعبر تمثيل الوثائق بنظام استرجاع المعلومات. ويُعد البتر Truncation في مصطلحات البحث أحد أبرز نماذج المضاهاة الجزئية. فعلى سبيل المثال عند البحث عن مصطلح *Information Technolog (يستخدم رمز النجمة للدلالة على البتر) فإن هذا الاستفسار سوف يسترجع وثائق تشتمل على Information Technolog, Information Technologist, Information Technologies كنتائج للمضاهاة الجزئية.

◄ 9.1.4 المضاهاة بالموضع
Positional Match

تتم المضاهاة بالموضع من خلال مراعاة موقع المعلومات بالوثائق أثناء عملية المضاهاة. ويُعد البحث التجاوري Proximity Searching نموذجاً لهذه الحالة. فإذا كان استفسار المستفيد هو متجر المواد المستعملة Used with Store فإن النتائج المسترجعة سوف تشتمل على وثائق تتضمن عبارات مثل:

Used	Book	Store
Used	Clothing	Store
Used	Furniture	Store

وتتم عملية المضاهاة هنا بين تمثيل الاستفسار وتمثيل الوثيقة فقط على الكلمة الأولى والكلمة الأخيرة، على أن تأتي بينهما أي كلمة أخرى، ويتم تجاهل الكلمة التي تأتي في الوسط أثناء عملية المضاهاة.

◄ 9.1.5 المضاهاة النطاقية
Rang Match

تنطبق المضاهاة النطاقية على العبارات الرقمية مثل البحث عن قيمة التخفيض Sale Amount أو التواريخ Dates أو العبارات ذات الترتيب الطبيعي مثل شهور السنة (يناير، فبراير،... ديسمبر) ويتم في المضاهاة النطاقية تحديد نطاق البحث بين نطاقين مثل الحد الأعلى Upper Limit للاستفسار مثل البحث عن الوثائق التي نشرت قبل عام 2002 والحد الأدنى Lower Limit مثل الوثائق التي نشرت بعد عام 1992 أو كليهما، مثل البحث عن الوثائق بين الفترة 1993 إلى 2002. من ثم فإن قواعد البيانات الرقمية وتواريخ النشر تُعد النماذج التقليدية البارزة للبحث النطاقي.

هـذه الأنـواع الأربعـة مـن نمـاذج المضاهـاة تتعامـل مـع الاستفسـار الأصلـي وتمثيـل الوثائـق دون الحاجـة إلـى إجـراء أي عمليـات حسـابية أو تغييـرات مثـل التـي تتـم علـى خوارزميـات البحـث. وعـادة مـا تظهـر مضاهـاة المصطلحـات فـي نمـوذج المنطـق البولينـي، أمـا فـي النمـاذج الأخـرى مثـل مسـاحة الزاويـة أو النمـوذج الاحتمالـي، فـإن مصطلحـات الاستفسـار وتمثيـل الوثائـق يتـم المضاهـاة بينهمـا بطـرق غيـر مباشـرة حيـث يتـم تحويلهـا إلـى مقاييـس تشـابه Similarity Measurement قبـل المضاهـاة بينهمـا.

9.1.6 مضاهاة مقياس التشابه

يمكـن إجراء مضاهـاة مقياس التشابه بطرق متنوعة. ففي نمـوذج الفراغ الاتجاهي علـى سـبيل المثـال تـتم المضاهـاة بالاعتمـاد علـى المسـافة بيـن الأسـهم أو درجـة مسـاحة الزاويـة Degree of Vector Angle فكلمـا كانـت مسـاحة الزاويـة صغيـرة، ازدادت درجـة التشـابه بيـن الاستفسـار والوثيقـة. وفـي النمـوذج الاحتمالـي يتـم حسـاب التشـابه علـى أسـاس تـردد المصطلحـات لتحديـد احتمـالات الصلاحيـة (العلاقـة) بيـن الاستفسـارات والوثائـق. ففـي هـذه النوعيـة مـن نظـم المضاهـاة، يتـم اختيـار مقيـاس تشـابه كمـي (المسـاحة، التـردد) وليـس المصطلحـات نفسـها، ويتـم إجـراء المضاهـاة النهائيـة بالاعتمـاد علـى هـذا المقيـاس الكمـي. وتجـدر الإشـارة إلـى أن مضاهـاة مقاييـس التشـابه تتيـح مـن ناحيـة أسـاليب إضافيـة وإجـراء عمليـات البحـث والاسـترجاع، إلا أنهـا مـن ناحيـة أخـرى يمكـن أن ينتـج عنهـا أخطـاء وضوضـاء وخاصـة فـي عمليـات حسـاب مقاييـس التشـابه ودرجاتهـا.

باختصـار وبصـرف النظـر عـن أسـلوب المضاهـاة، فـإن المضاهـاة هـي الآليـة الأساسـية لاسـترجاع المعلومـات. وسـوف تسـاعد النمـاذج التـي سـيتم مناقشـتها فـي بقيـة هـذا الفصـل فـي التعـرف إلـى كيـف تتـم عمليـات المضاهـاة فـي الظـروف المختلفـة، إلـى جانـب النمـاذج المختلفـة وملامحهـا ومزاياهـا وعيوبهـا.

9.2 نموذج المنطق البوليني

يرجع النموذج البوليني إلى مخترع فكرة المنطق البوليني جورج بولي George Boole والذي قدمه في منتصف القرن التاسع عشر. ويتعامل المنطق البوليني مع ثلاث معاملات منطقية أساسية هي:

- المعامل المنطقي للضرب Logical Product (X)
- المعامل المنطقي للجمع (+) Logical Sum
- المعامل المنطقي للفرق (ـ) Logical Difference

وفي مقابل تلك المعاملات المنطقية الثلاث تم توظيف المعاملات AND, OR, NOT لكي يتم استخدامها في العمليات المنطقية بنظم استرجاع المعلومات. وفي بدايات أنظمة استرجاع المعلومات على الإنترنت تم استخدام معامل الجمع (+) لتمثيل المعامل AND، ما أدى في بعض الأحيان إلى حدوث بعض الخلط لدى المستفيدين، لأنها تستخدم فعلياً (+) للدلالة على المعامل OR في دلالات المنطق البوليني.

- يعتمد المعامل AND على دمج مصطلحين أو أكثر معاً في عبارة البحث ويتطلب أن تظهر كل المصطلحات الواردة باستفسار المستفيد ويربطها المعامل AND بحيث تكون ممثلة في الوثيقة المسترجعة.

- يستخدم المعامل OR للجمع SUM حيث يقوم بالربط بين مفهومين أو مصطلحين مرتبطين بعلاقة ما معاً في عبارة البحث. ويستخدم للدلالة على ورود أي من تلك المصطلحات التي تحويها عبارة البحث المربوطة بالمعامل OR بالوثيقة المسترجعة أو كل أو بعض المصطلحات. من ثم فالوثيقة التي تشتمل على أي من المصطلحات التي تم تخصيصها في عبارة البحث يتم اعتبارها وثيقة صالحة ويسترجعها النظام.

- يساعد معامل الفرق أو المعامل NOT على تقييد البحث من خلال استبعاد المصطلحات الواردة بعد المعامل NOT من الاستفسار، من ثم استرجاع

الوثائق التي لا تشتمل على تلك المصطلحات واستبعاد الوثائق التي تشتمل عليها. وقد تم عرض العديد من النماذج والأمثلة على هذه الحالات المختلفة واستخداماتها في معالجات المنطق البوليني.

وكما سبقت الإشارة فإن مورتيمر تيوب Motimer Tupe هو أول من استخدم المنطق البوليني في استرجاع المعلومات. ومع تطور استخدام الأنظمة الآلية المحسبة في استرجاع المعلومات ازداد الاهتمام بتوظيف المنطق البوليني الذي أثبت جدارته وكفاءته في تمثيل التعبير عن استفسارات المستفيدين. وفي العصر الرقمي الذي يعتمد بصفة أساسية على استخدام الإنترنت في إتاحة المعلومات، يوجد عدد محدود جداً من الأنظمة التي لا تدعم النموذج البوليني في البحث والاسترجاع.

وقد أشار سبارك جونز وويليت (Spark Jones & Willet, 1997) إلى أن نموذج المنطق البوليني يعد أكثر الآليات انتشاراً وتطبيقاً في عمليات استرجاع المعلومات. لكن هذا لا يعني أن المنطق البوليني كنموذج لاسترجاع المعلومات يخلو من العيوب وأن كله مزايا، فعلى العكس من ذلك توجد العديد من الدراسات التي تناولت مقارنات مفصلة حول مزايا وعيوب نموذج المنطق البوليني عند تطبيقه باسترجاع المعلومات. ومن أمثلة هذه الدراسات (Chowdhury, 2010 ; Cooper,1988; Frants,et al,1999; Korfhage,1997; Spack - Jones &Willett,1997).

وسيتم فيما يلي استعراض تلك المزايا والعيوب بشيء من التفصيل:

9.2.1 مزايا نموذج المنطق البوليني

لقد أثبت التطبيق المكثف لنموذج المنطق البوليني باسترجاع المعلومات جدارة وكفاءة هذا النموذج بصورة واضحة. ويرجع ذلك لعدة أسباب:

أولاً: أنه يدعم معالجة الأوجه المتنوعة لاحتياجات المستفيدين، حيث يساعد على تفكيك الاستفسارات أو الوثائق إلى مفاهيم مستقلة والتعبير عن العلاقات بينها. فالمعامل AND يقوم بالدمج بين وجهين مختلفين، ما يساعد على التعبير عن الأوجه

المعقدة لاحتياجات المستفيدين وتضييق نطاق البحث، أما المعامل OR فيساعد على تحديد الأوجه المختلفة للاستفسار أو الوثيقة، ما يساعد على توسيع نطاق البحث من خلال توفير بدائل متنوعة للمصطلحات أو التعبير عنها بكلمات ذات علاقة مباشرة بها. ويساعد المعامل NOT على فصل الأوجه المعقدة إلى أوجه أكثر بساطة، من ثم يتمكن المستفيد من استبعاد الأوجه التي لا يرغب في ظهورها في قائمة النتائج النهائية. من ثم فإن تطبيق نموذج المنطق البوليني يساعد على تحقيق المرونة والفعالية لمستوى لا يمكن لأي نموذج آخر لاسترجاع المعلومات أن ينافسه فيه.

ثانياً: أن تطبيق نموذج المنطق البوليني بنظم استرجاع المعلومات أثبت فعالية كُلفته إلى المستوى الذي يجعله أحد المتطلبات الأساسية للمستفيدين من تلك النظم. فقد وصل عدد الأنظمة العالمية التي تطبق هذا النموذج في عمليات البحث والاسترجاع إلى الآلاف، حيث تُمكن هذه الأنظمة المستفيد النهائي من معالجة استفساره باستخدام معاملات المنطق البوليني لتوسيع أو تضييق أو حتى استبعاد بعض الأجزاء من المفاهيم. وذلك على الرغم من أن بعض الباحثين مثل (Belkin & Croft, 1987) يرون أن نموذج المنطق البوليني اكتسب شهرته من خلال الممارسة الواسعة وليس من خلال قوة نظريته.

ثالثاً: يتميز نموذج المنطق البوليني بسهولة فهمه (Spack - Jones &Willett,1997) وذلك على الرغم من أن عدداً محدوداً من الدراسات تناولت المقارنة بين ما يمكن للنظام تحقيقه في مقابل ما لا يستطيع أداءه كنموذج لاسترجاع المعلومات، والذي ربما يرجع إلى عاملين أساسين هما:

- **الأول**: أن نموذج المنطق البوليني هو الأقدم بين كل نماذج استرجاع المعلومات، ويعتقد الكثيرون أن مزاياه واضحة ولا تحتاج إلى تفسيرات إضافية.
- **الثاني**: أن نموذج المنطق البوليني تعرض للكثير من الانتقادات التي كان مصدرها أنه أقدم نموذج لاسترجاع المعلومات وعند ظهور أي نموذج جديد يتم تفنيد وانتقاد النموذج البوليني؛ فمن الطبيعي أن يقوم القائمون على

تطوير النماذج الجديدة بتحديد القيود التي توجد في النماذج الأقدم، ومنها النموذج البوليني، والتي يمكن للنموذج الجديد التغلب عليها. ومع ذلك فإن مصممي ومطوري نظم استرجاع المعلومات من ناحية والمستفيدين من ناحية أخرى يفضلون العمل مع نماذج يمكن فهمها بسهولة.

رابعاً: أن أنظمة استرجاع المعلومات القائمة على النموذج البوليني من السهل تطويرها عند مقارنتها بغيرها من الأنظمة، نظراً لأن الخوارزميات التي يعتمد عليها النموذج البوليني أكثر بساطة في التطبيق عن غيرها من الخوارزميات التي يتم تطبيقها في النماذج الأخرى.

ونتيجة لكل ما ذكر سابقاً من مزايا، تشمل طريقة المعالجة واتساع الاستخدام، فإن نموذج المنطق البوليني قد حظي باهتمام كبير في كل الدراسات التي تناولت نماذج استرجاع المعلومات.

9.2.2 صعوبات نموذج المنطق البوليني

سبقت الإشارة إلى أن قيود وعيوب نموذج المنطق البوليني تم دراستها وتناولها في العديد من الدراسات مثل: (Chowdhury, 1999; Cooper,1988; Frants et al., 1999; Korfhage,1997; Sparck Jones & Willett,1997) وسوف يتم استعراض أهم العيوب التي تناولتها تلك الدراسات فيما يلي:

أولاً: صعوبة التطبيق

من الصعب على أي مستفيد أن يستخدم المنطق البوليني في عمليات البحث والاسترجاع دون الحصول على القدر الكافي من التدريب والتأهيل والممارسة، وتكمن الصعوبة هنا في جانبين أساسيين هما:

- من الصعب على المستفيد اختيار المعامل البوليني الصحيح دون معرفة أو تدريب؛ حيث إنه عادة ما يحدث خلط لدى المستفيدين في معاني ودلالات المعاملين AND و OR نظراً لأن لكلا المعاملين معنى مختلف عن المعنى

التقليـدي المسـتخدم ودلالتـه الشـائعة، فالمعامـل AND عـادة مـا يسـتخدم فـي السـياق التقليـدي بمعنـى إضافـة (+) فعلـى سـبيل المثـال عنـد القـول إن المسـتفيد سـيجري بحثـاً فـي المحركيـن Google and Bing تعنـي أنـه سـيجري البحـث في كليهمـا. أمـا المعامـل OR فعـادة مـا يسـتخدم في السـياق العـام بمعنى أي منهمـا، فعنـد القـول إن الباحـث سـيجري بحثـاً فـي Google or Bing فذلـك يعنـي أنـه سـيجري البحـث فـي أي منهمـا، بمعنـى أنـه فـي السـياق العـام AND تعنـي البحـث فـي عـدد أكبـر مـن محـركات البحـث مـن OR وهـو عكـس مـا يتـم تطبيقـه فـي النمـوذج البولينـي. ويوجـد العديـد مـن المسـتفيدين الذيـن يفكـرون بنفـس المنطـق عنـد قيامهـم بإجـراء بحـث بولينـي؛ حيـث يسـتخدمون المعامـل AND عنـد رغبتهـم فـي البحـث عـن عـدد كبيـر مـن النتائـج، ويسـتخدمون OR لتضييـق نطـاق البحـث. ومـن الواضـح أن معامـلات المنطـق البولينـي لا تعمـل بهـذه الطريقـة، وقـد يـؤدي هـذا الخلـط بالمسـتفيدين إلـى اختيـار المعامـل الخطأ.

ومـن الملاحـظ أن المسـتفيد عـادة مـا يجـد صعوبـة فـي تركيـب المعامـلات البولينيـة وترتيبهـا بصـورة صحيحـة. فكمـا سـبقت الإشـارة إلـى أن البحـث البولينـي المركـب Compund Boolean Searching يتكـون مـن أكثـر مـن معامـل مـن المعامـلات البولينيـة، وأن الترتيـب الطبيعـي لمعالجـة المعامـلات البولينيـة هـو كالتالـي:

- تتم معالجة المعامل NOT أولاً.
- ثم يأتي المعامل AND ثانياً في الترتيب.
- وأخيراً تتم معالجة المعامل OR.

وفـي كثيـر مـن الأحيـان يمكـن اسـتخدام الأقـواس لتحديـد شـكل الترتيـب الطبيعـي لمعالجـة المعامـلات البولينيـة، وعـادة مـا يتـم ذلـك فـي العبـارات البحثيـة المعقـدة، والتـي تشـتمل علـى العديـد مـن العلاقـات. وقـد يختلـف الترتيـب فـي هـذه الحالـة عـن الترتيـب السـابق، نظـراً لأن الأقـواس فـي هـذا الحالـة تحـدد أولويـات المعالجـة عنـد التطبيـق. مـع العلـم أن هـذا الأسـلوب معقـد ونـادراً ما يسـتخدم في معالجـة الاحتياجـات البحثيـة المعقـدة،

ويتطلب هذا الأسلوب خبرة كبيرة في معالجة المعاملات البولينية وترتيبها والتركيب الاصطلاحي للمفاهيم التي تتضمنها العبارة البحثية. بالتالي فإن هذا الأسلوب لا يصلح للمبتدئين في عمليات البحث أو لغير المتخصصيين في أنظمة البحث والاسترجاع. فالتعامل مع القواعد الاصطناعية للترتيب مثل معالجة الأقواس الداخلية (TERM AND TERM OR TERM) ليس أمراً سهلاً على المستفيد المبتدئ، ويتطلب تدريباً وممارسة وإتقاناً لآليات التركيب الاصطلاحي والبوليني معاً.

ثانياً: صعوبة الاختزال لكل العلاقات بين المصطلحات في ثلاثة أشكال بولينية ثابتة

من الصعوبات التي تحد من إمكانيات النموذج البوليني عدم القدرة على التعبير عن العلاقات غير البولينية بين المصطلحات، مثل العلاقات العرضية Casual Relationship وذلك لعدم وجود معامل يحقق تلك النوعية من العلاقات في النموذج البوليني. نفترض أن أحد المستفيدين يبحث عن معلومات عن تطبيق الحاسب الآلي في التعليم Application of Computer in Education، فعند استخدام المعامل AND للربط بين المفاهيم المتنوعة وما ينتج عنها من استفسارات تكون العبارة البحثية في صورتها البسيطة كالتالي: **Computer and Education**

ومن الصعب أن يتم تمثيل المصطلح Application لأنه كلمة عامة مثل مقدمة Introducation ونظرة عامة Genral Overview.. الخ في بناء العبارة البحثية، ومن المفترض أن يتم التعبير عن هذه النوعية من المصطلحات بمعاملات تشملها، إلا أن النموذج البوليني قاصر عن توفير هذه النوعية من المعاملات التي تمكن المستفيد من تضمين هذه النوعية من المصطلحات في عملية البحث. لذلك تقتصر الصيغة البحثية على Computer AND Education مع ذلك فإن النتائج المسترجعة لهذه النوعية من الاستفسارات لن تقتصر فقط على معلومات عن استخدام الحاسب الآلي في التعليم، لكن ستشمل أيضاً معلومات عن تعليم الحاسب الآلي Computer Education وهو موضوع خارج نطاق اهتمام المستفيد في هذه الحالة، ما يجعل بعض النتائج المسترجعة تعالج مفاهيم ليس لها علاقة باحتياج المستفيد الأصلي

وتكون مضللة ومضيعة لوقت المستفيد الذي سيقضيه في فلترتها واستبعادها. وعلى ذلك فالنموذج البوليني يختزل كل العلاقات بين المفاهيم والمصطلحات في ثلاث معاملات بولينية يتم توظيفها للتعبير عن كل العلاقات والربط بين المفاهيم التي يتضمنها الاستفسار. من ثم يمكن القول إنه بصفة عامة كلما كانت العبارة البحثية معقدة، أدى ذلك إلى صعوبة تفسيرها وتمثيلها من خلال العلاقات البولينية، وذلك بسبب محدودية النموذج البوليني في التعبير عن العلاقات التي تخرج عن نطاق تلك العلاقات البولينية الثلاث.

ثالثاً: عدم القدرة على وزن المصطلحات

من القيود التي يفرضها النموذج البوليني في البحث والاسترجاع أنه لا يتيح إمكانيات لوزن المصطلحات أثناء البحث. ويرجع ذلك إلى عدم وجود آلية للوزن تمكن المستفيد من تحديد الأهمية النسبية للمفاهيم والمصطلحات التي يتضمنها الاستفسار، حيث يفترض النموذج البوليني أن كل المفاهيم أو المصطلحات الواردة في الاستفسار لها نفس الأهمية النسبية، وهو بالطبع أمر غير صحيح في معظم الأحيان. فعلى سبيل المثال، نفترض أن المستفيد يبحث عن موضوع إتاحة المعلومات والأمن Information Access AND Security وأن المستفيد يرغب في التركيز بصورة أكبر على موضوع الأمن، أو بعبارة أخرى أن المستفيد يرغب في الحصول على معلومات عن معالجة قضية الأمن في إتاحة المعلومات وليس معالجة الموضوعين بنفس الدرجة من الأهمية. فوفقاً للنموذج البوليني في استرجاع المعلومات لن تتحقق توقعات المستفيد لعدم وجود آلية لإعطاء وزن نسبي للمصطلحات عند البحث.

رابعاً: القصور في التعبير عن الصلاحية وترتيب النتائج

لا يتيح النموذج البوليني إمكانية التعبير عن الصلاحية الجزئية، حيث يقسم النموذج البوليني المواد إلى فئتين أساسيتين عند الاسترجاع هما:

- **صالحة**: أي يوجد مضاهاة تامة بين استفسار المستفيد وبديل الوثيقة (التسجيلة الببليوجرافية).

- **غير صالحة:** بمعنى عدم وجود مضاهاة بين استفسار المستفيد وبديل الوثيقة.

لذلك، فإن النموذج البوليني لا يتيح آلية لترتيب النتائج، ما يمكن المستفيد من تحديد أفضل 15 وثيقة مثلاً ضمن المواد المسترجعة مع ترتيبها وفقاً للأهمية النسبية. بالتالي يضطر المستفيد إلى فحص كل النتائج بنفس ترتيب استرجاعها والتي عادة ما تصل إلى بضعة آلاف. وتجدر الإشارة إلى أنه عادة ما يكون بعيداً عن الترتيب وفقاً للصلاحية النسبية ويستخدم نماذج عامة للفرز مثل الترتيب الهجائي أو الزمني. بالتالي لا يستطيع المستفيد التحكم في حجم المواد التي يفحصها وفقاً لمستوى الأهمية مقارنة بعدد النتائج المسترجعة.

خامساً: الصفرية في مقابل الفيضان

قد يحصل المستفيدون على نتائج صفرية Null Output أو فيضان من النتائج Output Overload عند إجراء البحث البوليني. وعادة ما تظهر النتائج الصفرية عندما يكون الاستفسار مقيداً بدرجة كبيرة. ويحدث ذلك عند الربط بين عدد من المصطلحات باستخدام المعامل AND. ومن ناحية أخرى يحدث فيضان النتائج عندما يكون الاستفسار عاماً وواسعاً بدرجة كبيرة. عادة ما يحدث فيضان النتائج عندما يتم الربط بين المصطلحات باستخدام المعامل OR. ويمكن للمستفيد في هذه الحالات أن يقوم بتعديل الاستفسار لزيادة أو تقليل عدد النتائج المسترجعة، إلا أن ذلك قد يؤدي إلى أن تكون النتائج المسترجعة غير مطابقة لما يبحث عنه المستفيد من البداية، وتقتصر فقط على نتائج الاستفسار المعدل.

وللتغلب على المشكلات والقيود التي يفرضها النموذج البوليني، اقترح كوبر (Cooper, 1988) بعض الحلول الممكنة مثل:

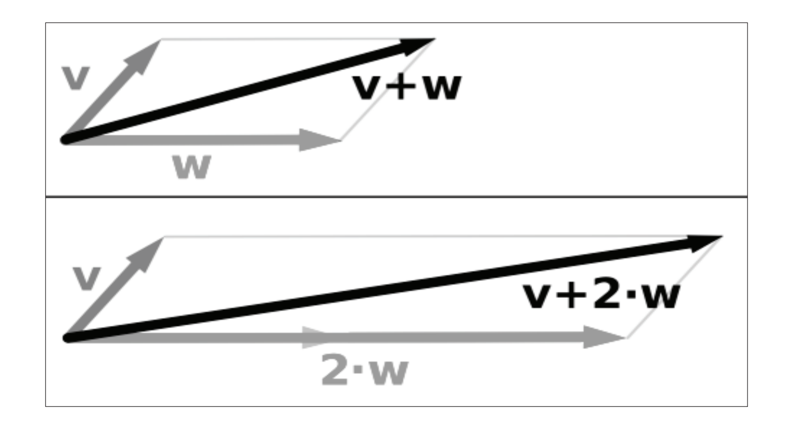
- إعداد استفسارات حرة خالية من المعاملات البولينية للتخلص من عيوب الاستفسارات البولينية. ومن الآليات المميزة لهذا المقترح تطبيق بعض الأنظمة لنماذج البحث Search Forms، ولم يحظ هذا المقترح بالقبول والتوسع في تطبيقه حتى منتصف التسعينات من القرن الماضي.

- كما تم تطوير عدد من الخوارزميات والنماذج الجديدة لاسترجاع المعلومات لتيسير عمليات ترتيب النتائج ووزن المصطلحات.. الخ. وعلى الرغم من كفاءة هذه النماذج من الناحية النظرية إلا أنها لم تحقق نجاحاً ملحوظاً عند تطبيقها في أنظمة استرجاع المعلومات البولينية (Korfhage, 1997, p. 63).

9.3 نموذج الفراغ الاتجاهي
Vector Space Model

يعد مجال الفراغ الاتجاهي أحد فروع علم الهندسة الفراغية والذي تم تطبيقه بكثافة في الجبر الخطي. ويشير إلى مجموعة من الأسهم التي يتم تجميعها لتكون فراغاً بحيث يمكن جمعها مع بعضها بعضاً وضربها بأعداد في هذا الفراغ. فعندما يتم تطبيق عمليات الجمع والضرب القياسي وبعض العمليات الأخرى على المتجهات (الأسهم) فإننا نصل لوصف كائن رياضي يطلق عليه فضاء اتجاهي.

يوضح المثال السابق نموذجاً لمعالجة مفهوم الفراغ الاتجاهي؛ فإذا كان لدينا ثلاثة أسهم يُطلق عليها متجهات تم تجميعها كما في الشكل، فإنه يمكن جمع وضرب الأسهم (المتجهات) في كميات قياسية للسهم v (باللون الأزرق) أُضيف



شكل (1/ 9) نموذج لتوزيع الموجهات في الفضاء الاتجاهي وطريقة قياسه (Sparck Jones & Willet, 1997)

إلى السهم w (باللون الأحمر، في أعلى الشكل)، وفي أسفله w ضُربت في معامل مساو لـ 2، ما أعطى المجموع v + 2*w.

وقد تم تطوير نموذج الفراغ الاتجاهي والذي يطلق عليه أيضاً معالجة المتجهات Vector Processing أو ناتج استرجاع المتجهات Vector Product Retreival على يد سالتون وزملائه Salton, et. al.، الذين قاموا ببناء نظام معالجة واسترجاع النصوص System for the Manipulation and Retrieval of Texts (SMART)، والذي تم توظيفه في سلسلة من بحوث وتجارب استرجاع المعلومات (Salton, 1968). وفي إطار عمليات تطبيق نظام SMART في بحوث ودراسات استرجاع المعلومات تم تطوير مجموعة من الآليات الجديدة في مجال استرجاع المعلومات في ذلك الوقت منها: وزن المصطلحات Term Weighting والمخرجات المرتبة Ranked Output.

ويعد نموذج الفراغ الاتجاهي النموذج الثاني من حيث أقدمية التطبيق ومن حيث الأهمية بعد النموذج البوليني في رحلة تطوير نماذج استرجاع المعلومات التي تعمل في البيئات التشغيلية (Sparck Jones & Willet, 1997).

ويتم التعبير عن كل مصطلح في نموذج الفراغ الاتجاهي على أنه بُعد Dimension، وعن الاستفسار على أنه مُتجه أو سهم Vector. ويتكون المُتجه من قيم أو درجات تعبر عن مجموعة المصطلحات المستخدمة في تمثيل الاستفسار أو الوثيقة، ويمكن أن تكون تلك القيم ثنائية Binary أو موزونة Weighted. في حالة القيم الثنائية يستخدم المعاملان (0.1) لتمثيل مدى ظهور المصطلح في المادة، وفي حالة القيم الموزونة تستخدم أرقام موجبة مثل (1.5, 0.3, 2.4, 5.9..etc). وتشير القيم الموزونة التي تستخدم للمصطلحات في الدلالة على الأهمية النسبية للمصطلح في تمثيل المادة (Kowalski, 2007). وقد حدد كروفهاج (Korfahge, 1997) طريقتين لوزن المصطلحات هما:

- خوارزميات موضوعية Objective لوزن المصطلحات مثل تردد المصطلحات أو حجم الوثيقة.
- خوارزميات غير موضوعية Subjective مثل استخدام أحكام المستفيدين

User Preception وقد سبق مناقشة العديد من طرق وزن المصطلحات والتي تعد قابلة للتطبيق من الناحية النظرية في نموذج الفراغ الاتجاهي.

- وتتميز كل خوارزمية من خوارزميات وزن المصطلحات بمجموعة من المزايا كما أن لكل منها عيوبها ومشكلاتها. وقد ناقش كورفهاج (Korfahge, 1997) بالتفصيل طرق التمثيل في كل من النوعين السابقين ومزايا وعيوب كل منهما عند تطبيقهما في تخصيص ووزن المصطلحات في المُتجه.

ويتم التعبير عن العلاقة في نموذج الفضاء الاتجاهي بأنه عبارة عن عدد الأبعاد Number of Dimension في الاستفسار أو الوثيقة والتي تعادل عدد المصطلحات المستخدمة في تمثيل المادة. وتتكون كل المتجهات (الأسهم) بالاستفسارات أو الوثائق من فضاء متعدد الاتجاهات. ويتم وصف موضع الاستفسار أو الوثيقة التي تمثله في الفضاء من خلال قياس إجمالي حزمة القيم المستخدمة في الدلالة على المصطلحات في المُتجه أو السهم (Sparck Jones & Willett, 1997).

ويتم تمثيل عملية إجراء البحث في نظم استرجاع المعلومات التي تعتمد على نموذج الفراغ الاتجاهي من خلال فحص المسافة، والتي تظهر في صورة متجه (سهم)، بين مُتجه الاستفسار والوثيقة في الفراغ الاتجاهي. ويتم في هذا النظام الحكم على درجة التشابه بين أي وثيقتين في النظام من خلال مقارنة درجة الأبعاد الممثلة ومن خلال حساب مقياس التشابه على أنه معامل التشابه أو الارتباط Cosine Coefficient. فإذا كان الاستفسار والوثيقة يعبران عن مفهوم متشابه فإن الزاوية التي بين الأسهم أو المتجهات تكون صغيرة، أما إذا كانا يتناولان مفهومين مختلفين فإن الزاوية بين الأسهم أو المتجهات تكون كبيرة (Lesk, 1997). من ثم يمكن بنفس الطريقة قياس التشابه بين الوثائق.

9.3.1 مزايا نموذج الفراغ الاتجاهي

أوضح سبارك جونز وويليت (Sparck Jones & Willett, 1997) المزايا التي يتمتع بها نموذج الفراغ الاتجاهي، وأنه يتيح لأنظمة استرجاع المعلومات أساساً

قوياً لعمليات التكشيف وتوظيف الصلاحية المرتدة Relevance Feedback وتصنيف الوثائق. فعند المقارنة بين نموذج الفراغ الاتجاهي والنموذج البوليني تتضح مزايا نموذج الفراغ الاتجاهي، وخاصة فيما يتعلق بالتغلب على جوانب القصور التي تمت مناقشاتها في نموذج المنطق البوليني. ويمكن إجمال هذه المزايا فيما يلي:

أولاً: إجراء البحث

لم يعد المستفيد بحاجة إلى فهم وتطبيق المعاملات البولينية المعقدة والتي تسبب له إرباكاً في كثير من الأحيان، عند إجراء البحث في نظم استرجاع المعلومات التي تعتمد على نموذج الفراغ الاتجاهي. فكل ما يحتاج إليه المستفيد عند التعامل مع نموذج الفراغ الاتجاهي هو اختيار مجموعة المصطلحات التي تلائم احتياجاته المعلوماتية بدقة عند إجراء البحث.

ثانياً: وزن المصطلحات

يتيح نموذج الفراغ الاتجاهي إمكانية وزن المصطلحات التي تعبر عن المفاهيم والمصطلحات التي تمثل الاستفسارات والوثائق، ما يساعد على تحديد الأهمية النسبية للمصطلح في الفراغ الذي يتم قياسه. فعلى سبيل المثال إذا كان لدى المستفيد استفسار عن أمن الشبكات Networks Security فإنه يستطيع أن يخصص وزناً أكبر للمصطلح أمن Security عن المصطلح شبكات Networks بالتالي لا تتم معالجة المصطلحين بالدرجة نفسها من الأهمية عند التكشيف والاسترجاع. من ثم فنموذج الفراغ الاتجاهي يتيح إمكانية تخصيص وزن للمصطلحات ما يساعد على تمثيل الاستفسار أو الوثيقة بدقة أكبر من حيث الأهمية النسبية للمعالجة التي يتناولها أي منهما.

ثالثاً: الترتيب

يتيح نموذج الفراغ الاتجاهي إمكانية ترتيب نتائج البحث ترتيباً تنازلياً وفقاً لصلاحية تلك النتائج لاستفسار المستفيد بحيث تأتي الوثائق الأكثر صلاحية على قمة قائمة النتائج المسترجعة. ويُعبر النموذج عن درجة التشابه Simialrity Score بين الوثائق والاستفسارات باستخدام مقياس درجات Scale من (0 إلى 1)، حيث تحصل الوثائق

الصالحة كلياً على درجة (1) ثم تحصل الوثائق الأقل صلاحية نسبياً على درجات 0.9, 0.8, 0.7.. إلخ وفقاً لمستوى صلاحية تلك الوثائق ودرجة تشابها مع الاستفسار. من ثم يمكن القول إنه في حين أن النموذج البوليني يستخدم مقياس صلاحية ثنائياً (صالحة أو غير صالحة) ما يعوق عمليات الترتيب والفرز وفقاً للصلاحية، فإن نموذج الفراغ الاتجاهي يتيح إمكانية ترتيب الوثائق بناء على درجة مقياس التشابه. بالتالي يتمكن المستفيد من تحديد الحد الأقصى من الوثائق التي يرغب في فحصها والاطلاع عليها من قائمة النتائج المسترجعة، بحيث يختار أفضل 10 وثائق وفقاً للترتيب والأهمية النسبية ويكون على يقين أن الوثائق الأخرى التي لم يفحصها هي أقل في الصلاحية من المجموعة التي قام بفحصها. وتجدر الإشارة إلى أن إمكانية تحديد عدد الوثائق التي يتم فحصها من مجموعة النتائج المسترجعة يعد تطوراً مهماً لخدمة المستفيدين من نظم استرجاع المعلومات التي تعتمد على هذا النموذج، حيث توفر تلك الميزة وقت وجهد المستفيد، نظراً لأنه لن يحتاج إلى استعراض وفحص كل الوثائق المسترجعة، كما هي الحال في النموذج البوليني، مع العلم أن عدد النتائج المسترجعة قد يصل إلى آلاف وأحياناً مئات الآلاف من الوثائق ما يتعذر معه فحصها بالكامل.

رابعاً: التغذية الراجعة للصلاحية Relevance Feedback

يعتمد نموذج الفراغ الاتجاهي على تطبيق مبدأ صلاحية التغذية الراجعة في تحسين أداء عمليات الاسترجاع، فبناء على صلاحية النتائج التي تم استرجاعها وعرضها مسبقاً، يقوم النظام بتخزين ردود أفعال المستفيدين عند التعامل مع نتائج البحث، ويستخدم تلك المعلومات المخزنة في تعديل صلاحية النتائج، ما يساعد على تحسين أداء الاسترجاع وعرض نتائج أكثر صلاحية بناء على تعاملات المستفيدين مع النظام. وتتم عمليات تخزين نتائج التغذية الراجعة للصلاحية دون تدخل من جانب المستفيد، وتتكرر تلك العملية أي عدد من المرات دون حد أدنى أو حد أقصى. وتظهر تلك الخاصية أو الميزة بوضوح في نظم استرجاع الإنترنت (محركات البحث) في خاصية نتائج مشابهة More Like This and More Similar Results.

ويتضح من العرض السابق أن نموذج الفضاء الاتجاهي يتميز بمجموعة من

الملامح ونقاط القوة التي تساعد في التغلب على مشكلات النموذج البوليني، إلا أن هذا النموذج لا يخلو أيضاً من بعض المشكلات التي تواجه أنظمة استرجاع المعلومات عند تطبيقه.

9.3.2 عيوب نموذج الفضاء الاتجاهي

يعتمد نموذج الفضاء الاتجاهي على مبدأ أساسي في بنائه هو إمكانية وزن المصطلحات من خلال حساب قيمتها في فضاء المصطلحات المستخدمة في النظام، ورغم جدارة هذا المبدأ والمزايا المتعددة التي يتمتع بها، إلا أنه أدى إلى بعض المشكلات في نموذج الفضاء الاتجاهي منها ما يلي:

أولاً: افتراض استقلالية المصطلحات

يفترض نموذج الفضاء الاتجاهي أن المصطلحات التي يتم اختيارها في عمليات التمثيل مستقلة عن بعضها البعض، وهذه الفرضية تعد من أهم عيوب هذا النموذج. فقد سبقت الإشارة إلى أن من أهم عيوب النموذج البوليني أنه لا يستطيع التعبير عن العلاقات خارج نطاق العلاقات البولينية. وقد كان من المتوقع أن يقوم نموذج الفضاء الاتجاهي بتوفير آليات أفضل للتعبير عن العلاقات، إلا أن الحقيقة أن هذا النموذج لا يوفر أي آلية للتعبير عن العلاقات بين المصطلحات بما فيها العلاقات البولينية. وبدلاً من حل مشكلة العلاقات القاصرة بالنموذج البوليني وضع فرضية أن المصطلحات التي يتم توظيفها باستفسارات المستفيدين لإجراء البحث بنظم استرجاع المعلومات التي تعتمد على نموذج الفضاء الاتجاهي مستقلة عن بعضها البعض ولا توجد علاقات تربط بينها.

ومن الواضح أن هذه الفرضية غير دقيقة وتفرض قيوداً غير عملية أثناء عمليات التمثيل والبحث. فإذا افترضنا أنه تم اختيار المصطلحات Automobile, Export, Import لوصف متجه Vector لوثيقة معينة، فهل يمكن افتراض أن هذه المصطلحات المستخدمة في تمثيل الوثيقة لا يوجد علاقات بينها. ولكن بالنظر إلى الواقع سنجد أن الوثيقة تتعامل مع Automobile Import, Automobile Export, Import and Export

and Automibile. ويعد افتراض استقلالية المصطلحات من الأهم الانتقادات التي وجهت إلى نموذج الفضاء الاتجاهي.

ثانياً: صعوبة تحديد المترادفات أو علاقات الجمل

من التحديات التي تواجه المستفيد عند استخدام نموذج الفضاء الاتجاهي هو التعبير بوضوح عن المترادفات أو علاقات الجمل بعضها ببعض، وذلك بسبب غياب المعاملات البولينية وتجاور المصطلحات. وبناء على ذلك لا يمكن استخدام المعامل OR لتحديد المترادفات مثل (Car OR Automobile) كما أنه لا يمكن تطبيق المعامل With لتكوين عبارات بحثية كما هي الحال في Information With Retrieval. مع العلم أن في عمليات البحث الحقيقية يحتاج المستفيد إلى التعبير عن المترادفات أو العبارات عند تمثيل الاستفسارات أو الوثائق. لذلك نجد أنه من الصعب إجراء البحث من دون المعاملات البولينية ومعاملات التجاور في أنظمة استرجاع المعلومات التي تعتمد على نموذج الفضاء الاتجاهي عندما يكون هناك حاجة إلى استخدام المترادفات وعلاقات الجمل في التعبير عن محتوى الاستفسارات أو الوثائق.

ثالثاً: عدم الموضوعية وتعقيد آليات الوزن

تعتمد أنظمة استرجاع المعلومات التي تستخدم نموذج الفضاء الاتجاهي على آليات معقدة وغير موضوعية لوزن المصطلحات. وتظهر عدم الموضوعية في عمليات وزن المصطلحات عندما يُطلب من المستفيد تخصيص وزن للمصطلحات وخاصة مصطلحات الاستفسار بناءً على رؤيته وأحكامه الشخصية. ويفترض هنا أن يقوم المستفيد بتقدير الأهمية النسبية للمصطلح الذي سوف يستخدمه وأن يحدد له وزناً نسبياً. بالتالي تظهر مشكلة عدم الموضوعية، حيث إن المستفيد في كثير من الأحيان يكون غير قادر على إعطاء وزن نسبي دقيق للمصطلح بالتالي تظهر مشكلة عدم الموضوعية. وعلى الجانب الآخر يتضح التعقيد في عمليات الوزن، حيث لا توجد خوارزمية خالية من العيوب وأوجه الانتقاد، كما أن الوصول إلى أفضل خوارزمية لبيئة استرجاع المعلومات يعد أمراً في غاية الصعوبة أن لم يكن مستحيلاً. فضلاً عن أن قواعد البيانات التي تبنيها أنظمة استرجاع المعلومات تتميز بالديناميكية الشديدة، حيث يتم تحديثها بصورة دائمة. بالتالي

فإن وزن المصطلحات لا بد أن يتغير ويتم تحديثه بصورة دائمة، لأن معاملات الوزن مثل تردد المصطلحات التي تطبقها خوارزميات الوزن تتغير مع تغير تركيبة قاعدة البيانات.

وقد قدم كوالسكي (Kowalski, 2007) عدداً من المسارات التي يمكن اتباعها لمعاجة قضية التغير الديناميكى بقواعد البيانات وتأثيره في خوارزميات الوزن، إلا أنه أشار إلى أن هذه المسارات سوف يكون لها تأثير واضح في عملية بناء وتطوير نظام استرجاع المعلومات من حيث التكلفة والوقت.

وبعيداً عن أوجه القصور الثلاثة التي تم ذكرها لنموذج الفضاء الاتجاهي، توجد بعض الصعوبات الأخرى مثل الحاجة إلى زيادة عدد المصطلحات المستخدمة في تمثيل الاستفسار حتى يتمكن المستفيد من صياغته بدقة، إلى جانب الحاجة إلى زيادة عدد المصطلحات المستخدمة في تمثيل الوثيقة أيضاً، ذلك حتى يتمكن النظام من التمييز الدقيق وتحسين أداء الاسترجاع. وذلك مقارنة بالنموذج البوليني الذي يمكن المستفيد من إجراء بحث جيد باستخدام عدد محدود من المصطلحات والربط بينها بالمعاملات البولينية. بالتالي ربما يكون استخدام عدد من اثنين إلى ثلاثة مصطلحات عدداً كافياً للتعبير عن الاستفسار أو تمثيل الوثيقة والحصول على نتائج ذات جودة عالية (Sparck Jones & Willett, 1997, p.259).

وتجدر الإشارة إلى أنه كلما زاد عدد المصطلحات التي يتم تعيينها للوثيقة أو الاستفسار ارتفعت التكلفة. كما أن هذا النموذج يفتقر إلى المبررات النظرية theoretical justification في بعض جوانب معالجة المتجهات (الأسهم) بالنموذج. فعلى سبيل المثال، اختيار مقياس معين لحساب درجات التشابه بين المتجهات لاستخدامه كنموذج لاسترجاع المعلومات لم يتم وصفه أو تبريره نظرياً، حيث تُرك تبريره للمستفيد (Slaton, 1989).

الشكل المثالي لهذا النموذج، أنه يضع الوثائق التي بينها علاقة صلاحية لاستفسار معين في مجموعة مترابطة في الفضاء، بينما تظهر الوثائق التي ليس بينها علاقة صلاحية منفصلة ومتباعدة في الفضاء (Salton, Wnag, Wnag, 1975). ومع ذلك فإن مضاهاة الاستفسار بمجموعة مترابطة من الوثائق، والتي يُطلق عليها مجموعة

الوثائق الافتراضية المجمعة Cluster Hypothetical Doucments أمر لم يكن من الممكن تحقيقه من دون تطبيقات هذا النموذج (Sparck Jones & Willett, 1997).

وقد بدأ تطبيق نموذج الفضاء الاتجاهي مع ظهور أنظمة استرجاع المعلومات على الإنترنت، ولم يتم تطبيقه فعلياً في أي بيئة استرجاع معلومات حقيقية قبل ظهور أنظمة الاسترجاع في بيئة الويب، حيث اقتصر تطبيقه قبل تلك الأنظمة على التجارب المعملية التي تمت على نظام SMART والذي ساعد على نمو ونضج هذا النموذج بصورة كبيرة، كما أن تطوير هذا النموذج ساعد على تطور البحوث والدراسات في مجال استرجاع المعلومات بصورة كبيرة.

9.4 النموذج الاحتمالي

Probablity Model

قام كل من مارون وكوهنز (Maron & Kuhns, 1960) بتطوير النموذج الاحتمالي لاسترجاع المعلومات في الستينيات من القرن الماضي، وقام كل من روبرتسون وسبارك بإجراء تطويرات إضافية على النموذج في السبعينيات (Robertson & Sparck, 1976). وقد أوضح كل من سبارك وويليت (Sparck Jones & Willett, 1997) أن الفكرة الأساسية التي يستند إليها النموذج الاحتمالي هي:

«تحاول نظم استرجاع المعلومات التي تعتمد على اللغة الطبيعية، والتي مازالت بعيدة عن الدقة، تحقيق معادلة التحديد المؤكد للوثائق الصالحة لاستفسار معين، وحيث أن هذا الوضع مضاد تماماً لعمليات الاسترجاع التي تحتاج إلى إزالة جميع جوانب الغموض لتحقيق هذه المعادلة عند البحث في وقواعد البيانات الرقمية (Sparck Jones & Willett, 1997, P.259) بالتالي فإنه إذا تم تطبيق نظرية الاحتمالات والتي يكون فيها الحدث له احتمالات تتراوح بدرجة نسبية بين صفر إلى 100، أو (0to1) (عند إجراء البحث».

بالتالي فإن هذا النموذج يراعي عناصر عدم اليقين Uncertainity Elements في معالجة عملية استرجاع المعلومات والتي تتمثل في: ما مستوى صلاحية وثيقة معينة تم استرجاعها لاستفسار معين؟ (Bookstein, 1985).

ويحاول النموذج قياس مدى احتمال صلاحية وثيقة معينة لاستفسار معين باستخدام مجموعة من الطرق الإحصائية التي يمكن من خلالها قياس الاحتمالات. ويطلق على هذه العملية في سياق استرجاع المعلومات احتمال الصلاحية The Probablity of Relevance بين استفسار ووثيقة.

وعلى خلاف غيره من نماذج استرجاع المعلومات فإن نموذج الاحتمالات لا يعالج الصلاحية على أنها مقياس مضاهاة أو عدم مضاهاة Miss- or- Match بل يعبر عن الصلاحية في إطار احتمالات ومستويات نسبية. فعلى سبيل المثال يقوم النظام بتحديد نسبة احتمال صلاحية وثيقة معية لاستفسار محدد، فيعرض مثلاً أن الوثيقة D قد تكون صالحة بنسبة 35 ٪ لاستفسار Q.

يعتمد النموذج الاحتمالي على طرق متنوعة لقياس الاحتمالات ومستويات الصلاحية النسبية بين الوثائق والاستفسارات من خلال حساب معدل التشابه بين الوثيقة والاستفسار. وتعتمد أحكام التشابه على أساليب قياس لعل أبرزها معدل تردد الكلمات Term Frequency. وبصفة عامة يمكن القول إنه في إطار هذا النموذج كلما ارتفعت درجة التشابه بين الاستفسار والوثيقة، زادت احتمالات صلاحية الوثيقة للاستفسار. ويتم في نظم استرجاع المعلومات التي تعتمد على النموذج الاحتمالي تحديد الوثائق التي يتم استرجاعها كنتائج للاستفسارات عندما تحقق تلك الوثائق فرضية أساسية تتمثل في أن تكون درجة احتمال تشابه تلك الوثائق أعلى من حد معين Specific Threshold في مستوى الصلاحية (Korfhage, 1997).

9.4.1 مزايا النموذج الاحتمالي

بالمقارنة بالنموذجين السابقين، البوليني والفراغ الاتجاهي، يتميز النموذج الاحتمالي بالمزايا التالية:

أولاً: يوفر النموذج الاحتمالي الأساس النظري للممارسات التي تم تطبيقها مسبقاً على أساس تجريبي مثل آليات وزن المصطلحات إلى جانب الإرشادات والإجراءات اللازمة لتطبيقها في عمليات استرجاع المعلومات (Salton, 1989, pp. 348-

349). فعادة ما توصف عمليات استرجاع المعلومات بأن لها مستويات عدم يقين Uncertainity متنوعة عند الحكم على علاقة الصلاحية بين الوثائق والاستفسارات، من ثم فإن استخدام مبدأ أحتمالات الصلاحية النسبية هو أكثر واقعية في التعبير عن صلاحية الوثائق وليس الصلاحية الثابته، إضافة إلى ذلك فإن العمليات الرئيسة الخاصة بهذا النموذج مثل قياس التشابه بين الوثيقة والاستفسار يتم تحديدها من خلال النموذج نفسه بدلاً من استخدام الأحكام الاعتباطية Herusitic Judegments، كما هو الحال في نموذج الفراغ الاتجاهي.

ثانياً: يفسر النموذج الاحتمالي مبدأ الاستقلالية في علاقات المصطلحات بالوثائق مثل علاقة ظهور وثيقة في عملية استرجاع المعلومات وتأثيره في ظهور وثيقة أخرى، حيث لم يعد المستفيدون بحاجة إلى افتراض الاستقلالية بين المصطلحات والذي يعد افتراضاً غير واقعي عند التطبيق كما هي الحال في نموذج الفراغ الاتجاهي. كما أن النموذج يوفر آليات لوزن المصطلحات وتحديد درجة التشابه النسبي بين الوثائق والاستفسارات ويُمكن المستفيد أيضاً من اختيار الوثائق الأكثر صلاحية.

ويتيح النموذج إمكانيات ترتيب النتائج المسترجعة وفقاً لصلاحيتها النسبية، حيث يفترض النموذج أن الوظيفة الأساسية لنظام استرجاع المعلومات هي مضاهاة الوثائق وتحديد درجة صلاحيتها من ثم ترتيبها ترتيباً تنازلياً وفقاً لاحتمالات الصلاحية المرتبطة باحتياجات المستفيدين (Sparck Jones & Willett, 1997) ويطلق على هذا الافتراض مبدأ الترتيب الاحتمالي Probablity Ranking Pribciple. ويساعد مبدأ الترتيب الاحتمالي على تمكين المستفيد من التحكم، إلى حد ما، في حجم النتائج المسترجعة من خلال التعبير عن الوزن والترتيب بصيغ احتمالية.

ثالثاً: استخدام معلومات التغذية الراجعة Relevance Feedback في تطوير طرق استرجاع أكثر كفاءة (Kowalski, 2007)، هذا إلى جانب قدرته على تحديد مواطن الضعف فيه بسهولة والعمل على تقويتها والتغلب عليها. يتميز النموذج الاحتمالي بإمكانية الطوير والتحسين الذاتي والذي يعد أحد أهم عناصر القوة في هذا النموذج.

رابعاً: النموذج الاحتمالي في شكله الأساسي لا يطبق المنطق البوليني الذي

يرى كثير من المستفيدين أنه آلية بحث صعبة التطبيق. مما يجعل من نظم استرجاع المعلومات التي تعتمد على النموذج الاحتمالي أكثر صداقة للمستفيد User Friendly من نظم استرجاع المعلومات التي تطبق المنطق البوليني.

9.4.2 عيوب النموذج الاحتمالي

تم تحديد عيوب النموذج الاحتمالي من أوجه متعددة منذ نشأته وعلى مر مراحل تطوره. ويمكن تلخيص هذه العيوب في العناصر التالية:

أولاً: الصلاحية الثنائية

على الرغم أن الصلاحية في النموذج الاحتمالي هي عبارة عن قيم متصلة تتراوح بين صفر وواحد، وليست قيماً ثنائية صفر أو واحد، كما هي الحال في النموذج البوليني، فإن النموذج الاحتمالي يفترض أن الصلاحية لها قيم ثنائية وهي كالتالي:

$$\Pr \ (nonrel) = \Pr \ (rel)$$

وتشير المعادلة إلى أن احتمال الصلاحية Pr (rel) تساوي احتمال عدم الصلاحية Pr (nonrel)

بمعنى آخر، أن قيم احتمال عدم الصلاحية ثابتة بمجرد حساب احتمال الصلاحية، أو الوثيقة لديها فرصتان هما أن تكون ضمن المجموعة الصالحة أو أن تكون ضمن المجموعة غير الصالحة. وذلك يلغي مبدأ عدم اليقين في عملية استرجاع المعلومات. وقد أوضح روبرتسون (Robertoson, 1976) أن القيم الثنائية لها مزايا متعددة، ألا أنها بالتأكيد ليست دقيقة بشكل عام أو كل الحالات.

ثانياً: تحسين نتائج الاسترجاع

لم تظهر فروق كبيرة في مستوى جودة النتائج المسترجعة من خلال النموذج الاحتمالي ، حيث لم يستطع تحسين كفاءة الاسترجاع بدرجة ملحوظة. فالنتائج التي يتم الحصول عليها من النموذج الاحتمالي رغم جودة عرضها، إلا أنها ليست أفضل من نتائج الاسترجاع في كل من النموذج البوليني ونموذج الفراغ الاتجاهي.

وهنا يظهر سؤال مهم هو: هل هناك حاجة إلى نماذج استرجاع معلومات جديدة في الوقت الذي تعمل فيه النماذج الحالية بدرجات متكافئة إلى حد كبير؟

وإلى جانب العيبين السابقين توجد بعض الأمور التي تحد من تطبيق هذا النموذج منها:

صعوبة التطبيق: وترجع صعوبة التطبيق إلى أنه نموذج معقد حسابياً ويتطلب عمليات حسابية مكثفة، مما يجعل فهمه نظرياً يحتاج إلى تطبيق آليات حسابية متنوعة تعتمد على نظرية الاحتمالات.

التنوع: يوجد للنموذج الاحتمالي أشكال متنوعة في المعالجات الحسابية ولا يوجد اتفاق بين المتخصصين على الطريقة المثلى للمعالجة الرياضية بين المهتمين به على الرغم من الاتفاق حول المبادئ الرئيسة للنموذج (Bookstein, 1985).

ندرة التطبيقات: كما هي الحال في نموذج الفراغ الاتجاهي فإن النموذج الاحتمالي لم يكن له تطبيقات حقيقية حتى ظهور نظم استرجاع المعلومات من الإنترنت، حيث اقتصر قبل ظهور تلك النظم على التجارب في البيئات المعملية.

9.5 التوسع في طرق استرجاع المعلومات

وضعت النماذج الثلاثة (البوليني والفراغ الاتجاهي والاحتمالي) التي تمت مناقشتها في هذا الفصل المنهجيات والقواعد الأساسية لاسترجاع المعلومات. ونتيجة لأثر تلك النماذج في البحث والتطبيق تم تطوير مجموعة من النماذج الجديدة التي توسعت للنماذج الثلاثة السابقة. فعلى سبيل المثال تم تطوير النموذج البوليني الموسع كامتداد لكل من النموذج البوليني ونموذج الفراغ الاتجاهي. كما تم وضع نموذج المجموعة الغامضة Fuzzy Set بالاعتماد على النموذج البوليني في بنيته الأساسية وباستخدام نظرية المجموعة The Set Theory وتطبيقها لأول مرة في مجال استرجاع المعلومات. كما أن نموذج تكشيف الدلالات الكامنة Latent Semantic Indexing مشتق من نموذج الفراغ الاتجاهي، كما تم تطوير نموذج شبكة الاستدلال Inference Network بالاعتماد على التوسع في النموذج الاحتمالي وآليات ترتيب النتائج ترتيباً احتمالياً تنازلياً بحيث تلبي احتياجات المستفيدين

بدلاً من احتمالية صلاحيتها لاحتياجاته والذي يعد أساس النموذج الاحتمالي (Sparck Jones & Willett, 1997). ويمكن التعرف إلى تفاصيل كاملة عن التوسعات التي جرت لنماذج استرجاع المعلومات في التقسيم الفئوي لنظم استرجاع المعلومات الذي قدمه (Baez – Yates & Ribeiro – Neto, 1999). وسوف تتم فيما يلي مناقشة اثنين من هذه التوسعات وهما النموذج البوليني الموسع ونموذج المجموعة الغامضة.

9.5.1 النموذج البوليني الموسع
Extended Boolean Model

سبقت الإشارة إلى أن من أهم عيوب النموذج البوليني عدم القدرة على وزن المصطلحات، كما أن من عيوب نموذج الفراغ الاتجاهي عدم توافر آلية للتعبير عن العلاقات البولينية. وللتغلب على هاتين المشكلتين اللتين تحدان من إمكانيات النموذجين تم تطوير النموذج البوليني الموسع لكي يوفر إمكانات لوزن المصطلحات والتعبير عن العلاقات البولينية، والذي يعد دمجاً بين مزايا النموذجين معاً. وتجدر الإشارة إلى أن العديد من الباحثين قاموا بالعديد من المحاولات لبناء هذا النموذج ومنهم بوكستين (Bookstein, 1978)، ويعد هارى أوو (Harry WU) أول من قدم مفهوم النموذج البوليني الموسع في رسالته للدكتورة التي كانت تحت اشراف جيرارد سالتون (Gerard Salton)، وقد استعرض فيها آليات عمل هذا النموذج والخوارزميات المقترحة لتنفيذه (Salton, Fox, & WU, 1983; WU, 1983).

ويتم في النموذج الموسع تخصيص وزن للمصطلحات باستخدام مزيج من المعاملات التالية:

- التقارب Proximity
- الموقع Location
- التردد Frequency
- الصلاحية المتوقعة Precieved Relevance

ويمكن من خلال هذا النموذج ترتيب النتائج بالاعتماد على إمكانات الوزن النسبي من ثم يمكن التحكم في عدد الوثائق التي يتم استرجاعها لكل استفسار. يضاف إلى ذلك المحافظة على إمكانات بناء العلاقات البولينية بين المصطلحات. وعلى الرغم من مزاياه السابقة إلا أن النموذج الوليني لم يتم تطبيقه بتوسع في أنظمة استرجاع المعلومات المستخدمة بقواعد البيانات الببليوجرافية للأسباب التالية:

أولاً: صعوبة تعيين وزن للمصطلحات بكفاءة ودقة بسبب العيوب نفسها التي تم ذكرها في نموذج الفراغ الاتجاهي.

ثانياً: فشل النموذج في استرجاع العدد نفسه من النتائج مع الاستفسارات المتساوية من ناحية بنية العلاقات البولينية عند تخصيص أوزان مختلفة لمصطلحات الاستفسار (Korfahge, 1997). فمن الطبيعي أن يتم استرجاع عدد أكبر من الوثائق للمصطلحات التي لها وزن نسبي مرتفع والذي يراه البعض نتيجة غير منطقية حيث إن عدد الوثائق الصالحة ثابت ويجب ألا يتغير وما يتغير هو ترتيبها وفقاً للوزن النسبي للمصطلحات.

ومع ذلك فإن النموذج البوليني الموسع دمج بين مزايا النموذج البوليني ونموذج الفراغ الاتجاهي فساعد العديد من محركات البحث على الاستفادة من مزايا النموذجين، وقامت العديد من المحركات بتطبيقه في بحث الإنترنت ومنها محرك البحث جوجل.

9.5.2 نموذج المجموعة الضبابية

Fuzzy Set Model

يعد الأذربيجاني لطفي زاده أول من قدم هذا النموذج في مجال استرجاع المعلومات في عام 1965 (Zadeh, 1965) بغرض التغلب على عيوب النموذج البوليني من خلال استخدام آليات التعبير عن الصلاحية الجزئية Partial Relevancy لنتائج البحث وذلك من خلال تطبيق مبادئ نظرية المجموعة Set Theroy.

في هذه النظرية يتم التعبير عن المادة على أنها إما ضمن مجموعة أو ليست ضمن

مجموعة، كما أن الوثيقة إما أن تكون صالحة أو غير صالحة لاستفسار معين، كما هي الحال في النموذج البوليني. ويساعد ذلك على وضع حدود فاصلة بين أعضاء المجموعة وغير الأعضاء بالمجموعة أو الوثائق الصالحة والوثائق غير الصالحة. إلا أن هذا الخط الحاد الفاصل بين الوثائق الصاحة وغير الصالحة غير موجود فعلياً في مجال استرجاع المعلومات، نظراً لأن الأنظمة وغالباً المستفيدين لا يمكنهما بدقة تحديد ما إذا كانت الوثيقة صالحة لاستفسار معين أم لا (Korfahge, 1997). لذلك تعد الصلاحية الجزئية انعكاساً أو تعبيراً أكثر دقة للتغلب على هذه المشكلة وإصدار أحكام واقعية.

وقد أطلق على الصلاحية الجزئية التي تم تطبيقها لتحسين إمكانيات النموذج البوليني نظرية المجموعة الضبابية. ويفترض هذا النموذج أن الوثائق والاستفسارات الضبابية هي الأساس في استرجاع المعلومات لذلك لا بد من وضع آلية لإصدار أحكام ضبابية بشأنها. ويعتمد هذا النموذج على تحديد مدى عضوية المادة ضمن المجموعة في مدى بين الدرجتين صفر إلى واحد، حيث يشير واحد إلى العضوية الكاملة، وتشير أي درجة بين الواحد والصفر إلى العضوية الجزئية. لذلك فالحدود التي تفصل بين الأعضاء وغير الأعضاء تصبح ضبابية ويحددها مستوى ودرجة العلاقة داخل المجموعة.

فعلى سبيل المثال يمكن تحديد مجموعة الطلاب المتميزين من بين كل الطلاب بطريقتين أساسيتين هما:

الأولى: تطبيق الطريقة التقليدية والتي يتم فيها تحديد مجموعة الطلبة الأوئل الذين حصلوا على أعلى متوسط درجات من بين المجموعة الكاملة، فمثلاً يتم تحديد الطلاب الذي حصلوا على متوسط أعلى من 3.9 كمتوسط درجات، وأي طالب يحقق هذه الدرجة يدخل ضمن مجموعة المكرمين، في حين أن أي طالب يحصل على درجة أقل من 3.9 فلن يكون ضمن مجموعة المكرمين. ذلك على الرغم أن بعض الطلاب قد حصلوا على متوسط درجات 3.89 والفرق بينهم وبين المجموعة الأولى غير ملحوظ.

الثانية: تعتمد على تحديد طلاب المجموعة على أساس الدرجة التي يحصلون

عليها، فالطلبة الذين يحصلون على درجة 3.9 أو أكثر مثلاً يحصلون على عضوية كاملة تعادل الدرجة (1 0.9)، بينما يحظى الطلبة الذين يحصلون على درجة أقل من 3.5 - 3.9 بعضوية جزئية، والمجموعة التي تحصل على درجة أقل من 3.5 على عضوية قريبة من الصفر، من ثم يتحدد مستوى العضوية بناء على مدى قربه أو بعده من الدرجة 1.0، بحيث يحظى الطالب الذي حصل على درجة 3.8 مثلاً بعضوية تعادل 0.8 بالتالي ويستبعد الطلبة الذين حصلوا على عضوية أقل من 3.5 وفقاً للمستوى الذي تم تحديده لمدى العضوية.

وعند تطبيق نظرية المجموعة الضبابية في استرجاع المعلومات فإن حكم الصلاحية على الوثيقة لا يعتمد على مقياس ثنائي بأن الوثيقة صالحة أو غير صالحة، كما هي الحال في النموذج البوليني. فبدلاً من تطبيق مقياس ثنائي يتم تطبيق مستوى عضوية لمجموعة الوثائق على أساس مدى قرب الوثيقة من مستوى الصلاحية. ويتم تحديد مستوى صلاحية الوثيقة بالمجموعة الضبابية أثناء عملية التكشيف (Bookstein, 1985).

ومن أهم مزايا نموذج المجموعة الضبابية أنه يتيح إمكانية تحديد مستويات صلاحية للوثائق، بحيث يتيح الوصول إلى الوثائق ذات الصلاحية الجزئية، مما يتيح للنموذج ترتيب النتائج ترتيباً تنازلياً وفقاً لمدى عضويتها بالمجموعة، ومستوى صلاحيتها. بالتالي يتمكن المستفيد من اختيار وعرض النتائج الأكثر صلاحية والتي تظهر في قمة قائمة النتائج. إضافة إلى ذلك يحافظ هذا النموذج على إمكانية بناء العلاقات البولينية بين المصطلحات. بالتالي تتميز نظم استرجاع المعلومات التي تطبق نموذج المجموعة الضبابية بإمكانيات الاسترجاع الاكتشافي Discovery Retrieval.

ومع ذلك لا يتيح نموذج المجموعة الضبابية المرونة الكافية التي تسمح بتعيين وزن لمصطلحات الاستفسار في مقابل مصطلحات الوثيقة، حيث تعتمد درجة استرجاع الوثيقة على الدرجة التي تحصل عليها أثناء التكشيف فقط، ولا يراعي مصطلحات الاستفسار (Salton, 1989). وتتضح عدم المرونة في نموذج المجموعة الضبابية عند التعامل مع العلاقات البولينية وعدم وزن مصطلحات الاستفسار عند تطبيق المعامل OR للتعبير عن العلاقة بين ثلاثة مصطلحات (A OR B OR C)

فالنموذج في هذه الحالة سوف يسترجع الوثائق D1, D2.. الخ، ويعطي الوثيقة D1 التي تشتمل على المصطلح A فقط الدرجة نفسها التي تحصل عليها الوثيقة D2 التي تشتمل على المصطلحات الثلاثة A OR B OR C وذلك لعدم قدرة النموذج على وزن مصطلحات الاستفسار. ومن الواضح في هذه الحالة أن درجة صلاحية الوثيقة D1 تم الحكم عليها من مصطلح واحد فقط هو المصطلح A كنتيجة لحقيقة أن مصطلحات الاستفسار لا يتم وزنها في هذا النموذج.

كذلك الحال عند تطبيق المعامل AND، فعند البحث عن المصطلحات (A AND B AND c) فإن الوثيقة D1 التي تشتمل على المصطلحين A AND B لن يسترجعها النظام لأنه سيعتبرها وثيقة غير صالحة، كذلك الحال بالنسبة للوثيقة D2 التي تشتمل على المصطلح A فقط أو الوثائق التي تشتمل على مترادفات لهذه المصطلحات. إضافة إلى ذلك فإنه عند مقارنة نموذج المجموعة الضبابية بنموذج الفراغ الاتجاهي، فإن نموذج المجموعة الضبابية لا يتيح أي آلية لتوسيع الاستفسارات. وعلى عكس النموذج الاحتمالي فإن نموذج المجموعة الضبابية ليس بمستوى النموذج الاحتمالي من ناحية قوة الأساس النظري، لذلك لم يحظ هذا النموذج بتطبيقات كاملة وقد تم تطبيقه بصورة متقطعة في بعض النظم المحدودة لأغراض التجربة والاختبار.

9.6 نماذج أخرى لاسترجاع المعلومات

تمت الإشارة في بداية هذا الفصل إلى أن نماذج الاسترجاع التي تم استعراضها هي وامتداداتها كلها نماذج تم تطبيقها في أنظمة استرجاع معلومات بصورة أو بأخرى، وإضافة إلى هذه النماذج توجد مجموعة أخرى من نماذج استرجاع المعلومات التي تم تطويرها تعتمد على آليات التفاعل بين المستفيد والنظام ولعل أبرزها مجموعة النماذج المعرفية Cognitive Models الذي يعتمد على العوامل الخاصة بالمستفيد User Factors في استرجاع المعلومات. وقد تمت الإشارة إلى أن هذه النوعية من النماذج لن يتم مناقشاتها في هذا الكتاب. وسيتم فيما يلي عرض ملخص عام للملامح الرئيسة للنماذج الثلاثة التي تم استعراضها في هذا الفصل.

9.7 ملخص عام لنماذج استرجاع المعلومات

يستعرض الجدول التالي الملامح الرئيسة للنماذج الثلاثة حيث يقارن بين تلك النماذج من خمس زوايا أساسية هي:

1. دعم المنطق البوليني
2. التعامل مع وزن المصطلحات
3. دعم ترتيب النتائج
4. معايير المضاهاة المطبقة بالنموذج (تحديد مدى التشابه بين الاستفسارات والوثائق).
5. ملامح إضافية مميزة.

ومن الملاحظ أن هذه الملامح الخمسة تعبر بشكل عام عن معايير الحكم على نقاط القوة والضعف في نماذج استرجاع المعلومات. فعلى سبيل المثال، يشير الملمح الخاص بدعم النموذج للمنطق البوليني إلى قدرة النظام وتمكين المستفيد من هيكلة الاستفسارات وبناء العلاقات بين المصطلحات. وعلى الجانب الآخر للميزة نفسها والمتعلقة بدعم المنطق البوليني فإنها تؤدي إلى فقدان النظام لميزة سهولة الاستخدام بسبب صعوبة تركيب العلاقات عند إجراء البحث البوليني.

جدول (9.1) يلخص النماذج العامة لاسترجاع المعلومات ومزاياها وعيوبها:

الملامح / النماذج	المنطق البوليني	الفراغ الاتجاهي	الاحتمالي
المنطق البوليني	نعم		
الوزن		نعم	نعم
الترتيب		نعم	نعم

معايير المضاهاة	ظهور المصطلحات	مساحة التوجيه (السهم الموجه)	تردد المصطلحات
ملامح إضافية مميزة		الصلاحية الراجعة	

ويتضح من الجدول أن النموذج البوليني هو الأضعف بين النماذج الثلاثة من حيث المزايا، فالنموذج البوليني يدعم فقط البحث البوليني، وتتم المضاهاة بناء على استخدام المصطلح الذي يبحث عنه المستفيد بوثائق النظام أو عدم استخدامه. مع ذلك فإن النموذج البوليني هو أكثر نماذج استرجاع المعلومات تطبيقاً في أنظمة قواعد البيانات الببليوجرافية على وجه الخصوص. أما النموذجان الآخران فيبدو أنهما سطحياً متشابهان، من حيث العمل على تطبيق آليات لوزن المصطلحات وترتيب النتائج وعدم تطبيق آليات البحث البوليني. ويختلف النموذجان فيما بينهما في معايير وزن المصطلحات وترتيب النتائج. إضافة إلى ذلك تميز نموذج الفضاء الاتجاهي باستخدام آليات الصلاحية الراجعة كملمح فريد من ملامح الأنظمة المطبقة لهذا النموذج. وقد بُذلت جهود كبيرة لبناء أنظمة تطبق آليات وزن المصطلحات وترتيب النتائج بالاعتماد على النموذجين (الفضاء الاتجاهي والاحتمالي)، بحيث تتيح إمكانيات أكثر فعالية وكفاءة من النموذج البوليني، إلا أن هذه الأنظمة لم تستطع تحقيق تميز ملحوظ في أدائها الاسترجاعي عن النظم التي تعتمد على نموذج المنطق البوليني (Korfahge, 1997).

9.8 العلاقة بين نماذج استرجاع المعلومات وآليات الاسترجاع

تم في الفصل الخامس مناقشة واستعراض آليات البحث والاسترجاع المختلفة، ومن الضروري التعرف إلى العلاقة بين نماذج استرجاع المعلومات وآليات الاسترجاع التي تمت مناقشتها. فعلى الرغم من عدم وجود علاقة واحد لواحد One to One بين كل منها، إلا أن بعض آليات الاسترجاع ترتبط بوضوح بنماذج استرجاع المعلومات التي اشتقت منها. فعلى سبيل المثال يرتبط البحث البوليني بنموذج المنطق البوليني بشكل

مباشر حيث إنه تطبيق واضح المعالم لهذا النموذج، كما أن توسيع الاستفسارات وخاصة باستخدام آليات الصلاحية الراجعة يرتبط بشكل مباشر بنموذج الفراغ الاتجاهي، كما أن البحث بالوزن يعتمد على خوارزميات تم تطويرها بالاعتماد على النموذج الاحتمالي وغيرها من نماذج الاسترجاع مثل النموذج البوليني الموسع.

وعلى الجانب الآخر توجد آليات استرجاع أخرى اعتمدت على نماذج استرجاع المعلومات الإضافية، فعلى سبيل المثال اعتمد البحث التجاوري في جذوره الأساسية على البحث البوليني الموسع. وعلاوة على ذلك تم تطبيق بعض آليات استرجاع المعلومات في أنظمة لم يتم تصميمها بالاعتماد على النموذج الذي اشتقت منه هذه الآليات، حيث تم تطبيقها جنباً إلى جنب مع آليات تلك النماذج بصرف النظر عن مصدرها، بالتالي فإن تصميم النظام يعتمد على تطبيق آليات استرجاع أكثر من تطبيقه لنماذج استرجاع. وتخلط النظم في كثير من الأحيان بين أكثر نموذج بغرض تطبيق آليات استرجاع متنوعة. لذلك تظهر الحاجة إلى تطوير نظم متعددة النماذج لاسترجاع المعلومات Multimodel IR System.

فالمعرفة الدقيقة للعلاقة بين نماذج استرجاع المعلومات وآليات الاسترجاع تساعد على اختيار النظام الملائم وفقاً للمهام التي يجب أن تؤديها تلك النظم. فعلى سبيل المثال لا بد من تطبيق النموذج البوليني في حالة حاجة المستفيد إلى إجراء بحث بوليني، أما في حالة حاجة المستفيد إلى وزن المصطلحات البحثية وترتيب النتائج هنا تظهر الحاجة إلى نموذج الفراغ الاتجاهي أو النموذج الاحتمالي ويتم تحديد أيهما الأنسب بناء على احتياجات المستفيدين من النظام أيضاً.

9.9 نحو نظم استرجاع معلومات متعددة النماذج
Multimodel IR Systems

لكل نموذج من نماذج استرجاع المعلومات التي تم استعراضها في هذا الفصل مزاياه وعيوبه، من ثم فإن النظم التي تطبق هذا النموذج سوف تؤدي وظائف استرجاع معينة وفقاً لإمكانيات هذا النموذج. لذلك من الضروري أن يعمل نظام

استرجاع المعلومات على الافادة من المزايا التي تتمتع بها كل النماذج من خلال دمج تلك النماذج في نظام متعدد النماذج. وقد تم التعبير عن المفهوم نفسه في دراسات فرانتس وآخرون (Frants, et, el., 1999) حيث أطلقوا على هذه النوعية من الأنظمة مصطلح أنظمة متعددة الإصدارات Mutiversion Systems. وتشير الممارسات الحالية في أنظمة استرجاع المعلومات إلى أن النموذج البوليني هو النموذج الأكثر انتشاراً وتطبيقاً في أنظمة استرجاع المعلومات البليوجرافية. ويتم تطبيق النماذج الأخرى تدريجياً في أنظمة استرجاع المعلومات على الإنترنت. فإذا كانت استفسارات المستفيدين تتراوح بين استفسارات بسيطة ومحدودة من حيث التعقيد إلى استفسارات مركبة ومعقدة بدرجة كبيرة، لا بد من أن يكون تصميم نظام استرجاع المعلومات قادر على التكيف مع تلك الاحتياجات المتنوعة من خلال تطبيق النظم متعددة النماذج. وتتطور أنظمة استرجاع المعلومات متعددة النماذج مع تطور أنظمة وآليات البحث على الإنترنت والتي أصبحت المنصة الرئيسة للوصول إلى المعلومات في العصر الرقمي.

ويوجد العديد من الأسئلة التي مازالت مطروحة وتظهر بشكل متوالٍ عن كيفية تطوير الأنظمة متعددة النماذج من خلال الدراسات والتجارب التي تتم في مؤتمرات استرجاع المعلومات مثل مؤتمر TREC وغيره من المؤتمرات التي تقدم إرشادات وتوجيهات عن كيفية بناء النظم الحديثة في هذا الجانب وضرورة إجراء دراسات مسحية للمستفيدين للتعرف إلى احتياجاتهم المعلوماتية وأساليب البحث التي يفضلونها في العصر الرقمي.

المصادر

- Baeza-Yates, R., Ricardo and Ribeiro-Neto, Berthier (1999). Modern Information Retrieval. New York
- ACM Book Press.
- Blekin, Nickolas. J., and Croft, W, Bruce (1987). Retrieval Techniques" Annual Review of Information Science and Technology, 22, 109-145.
- Bookstein, A. (1978). On the perils of merging Boolean and weighted retrieval systems.Journal of the American Society for Information Science,29(3), 156-158.
- Bookstein, A. (1985). Probability and fuzzy-set applications to information retrieval. Annual review of information science and technology, 20, 117-151.
- Chowdhury, G. G. (2010). Introduction to modern information retrieval. Facet publishing.
- Cooper, W. S. (1988). Getting beyond Boole. Information Processing & Management, 24(3), 243-248.
- Frants, V. I., Shapiro, J., Taksa, I., & Voiskunskii, V. G. (1999). Boolean search: Current state and perspectives. Journal of the American Society for Information Science, 50(1), 86-95.
- Ingwersen, P., & Järvelin, K. (2006). The turn: Integration of information seeking and retrieval in context (Vol. 18). Springer Science & Business Media.
- Kowalski, G. J. (2007). Information retrieval systems: theory and implementation (Vol. 1). Springer.
- Korfhage, R. R. (1997). Information Retrieval and Storage. New Yowrk: John Willy and Sons
- Lesk, Micheal (1997) Practical digital library: books, bytes, and bucks. San Francisco, California: Morgan Kaufmann, p. 297
- Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. Journal of the ACM (JACM), 7(3), 216-244.
- Robertson, S., & Sparck—Jones, K. (1976). Relevance weighting of search term. Journal of American Society for Information Science, 1(3), 129 -146.
- Salton, G. (1968). Automatic information organization and retrieval. New York: MaGraw-Hill

- Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. Reading: New York: Addison-Wesley.
- Salton, G., Fox, E. A., & Wu, H. (1982). Extended Boolean information retrieval. Cornell University.
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.
- Sparck Jones; Karen and Willet Peter (Ed.). (1997). Readings in information retrieval. San Francisco: Morgan Kaufmann.
- Soukhanov, Anne H., et. Al. (Eds.) (1984). Webster's, I. I. New Riverside University Dictionary, Boston: Riverside Publishing Co.
- Zadeh, L. A. (1965). Fuzzy Sets. Information and Control. 8(3), 338-353.

الفصل العاشر

تمثيل المعرفة على الإنترنت

مقدمة

لقد أدى تطور ونمو الشبكة العنكبوتية (WWW or The Web) إلى حدوث تغيير كبير في أساليب البحث عن المعلومات وسبل الإفادة من المصادر المتاحة من خلال شبكة الإنترنت. ويرجع ذلك بشكل كبير إلى النمو السريع والهائل في عدد وأشكال وأنواع مصادر المعلومات المتاحة من خلال الشبكة العنكبوتية، إضافة إلى تنوع تلك المصادر، وسهولة الوصول إليها، هذا إلى جانب طبيعة تلك المصادر والتكنولوجيات المستخدمة في إتاحتها. وقد جعلت هذه التطورات من الشبكة العنكبوتية أكبر مصدر للمعلومات في العصر الحالي (Bokor, 1999). وقد صاحب ذلك تنوع في أساليب استرجاع المعلومات المتاحة من خلال بيئة الويب. ونستعرض فيما يلي تطور أدوات استرجاع المعلومات في بيئة الويب.

10 نشأة أدوات الوصول إلى المعلومات في بيئة الويب وتطورها

قام عالم الفيزياء تيم برنر لي بوضع أسس الشبكة العنكبوتية في بداية التسعينيات من القرن العشرين لتكون وسيلة أساسية للباحثين في تبادل مسودات البحوث والرسائل الإلكترونية. ومنذ ذلك التاريخ بدأت العديد من الجامعات استخدام هذه الأداة في بث وتيسير سبل الوصول إلى المعلومات. ومع بداية عام 1993 كان هناك بضع مئات من المواقع المتاحة على الشبكة العنكبوتية معظمها مواقع تتعلق بكليات ومعاهد بحثية. وكانت الطريقة الأساسية لتبادل المعلومات بين مستخدمي الشبكة العنكبوتية في ذلك الوقت تتم من خلال بروتوكول تبادل الملفات المعروف بـ (File

Transfer Protocol (FTP وهو عبارة عن برنامج يمكن من خلاله نقل الملفات من حاسب إلى حاسب آخر من خلال واجهة تعامل تعمل بالأوامر. في تلك المرحلة إذا أراد شخص أن يسترجع معلومات من الشبكة العنكبوتية فعليه أن يتعامل معها من خلال هذا البروتوكول. وكانت هذه الطريقة فعالة في ظل مجموعات الويب الصغيرة، ولكن مع تزايد المجموعات ونموها لم تصبح هذه الوسيلة فعالة بالدرجة الكافية، مما دفع الباحثين للتنقيب عن وسائل أخرى. وتمثل أول تلك الحلول في الاعتماد على أحد محركات البحث التي تم تطويرها قبل نشأة الشبكة العنكبوتية والذي عُرف بالأرشيف Archive إلا أن استخدامه من خلال نظام التشغيل يونكس UNIX فرض ضرورة اختصار الاسم إلى Archie. وقد قام بتطوير هذا المحرك أحد طلاب جامعة ماكجيل McGill بمدينة مونتريال الكندية اسمه ألن إمتاج Alan Emtage. وقد اعتمد هذا المحرك أساساً على قاعدة بيانات بأسماء الملفات المتاحة على الشبكة العنكبوتية، فكانت عملية المضاهاة تعتمد بشكل كبير على البحث في قاعدة البيانات عن اسم الملف الذي يرغب المستفيد في استرجاعه (Gromov, 2000). وقد مرت عملية بناء وتطوير أدوات الاسترجاع في بيئة الويب بأجيال متعددة نذكر منها ما يلي:

- **الجيل الأول**

في عام 1993 طورت جماعة الاهتمام بالحاسبات بجامعة نفادا بالولايات المتحدة محرك بحث جديداً اعتمد على البنية نفسها المستخدمة في المحرك Archie وعُرف هذا المحرك الجديد بـ Veronica. والاختلاف الوحيد بين Archie و Veronica هو أن الثاني كان يعمل مع ملفات النصوص Plain Text Files، بينما كان الأول يعمل فقط على الاسترجاع من قاعدة بيانات تشتمل على أسماء الملفات. ثم ظهر تقريباً في التاريخ نفسه محرك ثالث عُرف بـ Jughead وقد اعتمد أيضاً على البنية نفسها المستخدمة في المحرك Veronica، وقد تم استخدام كل من Jughead and Veronica لتبادل الملفات من خلال أداة التصفح جوفر Gopher والتي قام بتطويرها مارك ماكهيل Mark McCahill في جامعة ميناسوتا لكي تحل محل المحرك Archie Lensse, (2004).

وفـي عـام 1993 ظهـر أول روبـوت[1] علـى يـد ماتثـوي جـاري Matthew Gary والـذي عُـرف بمتجـول الشـبكة العنكبوتيـة WWW Wanderer. وقـد كان الهـدف الأسـاسـي مـن هـذا الروبـوت هـو إحصـاء معـدل الزيـادة فـي الشـبكة العنكبوتيـة مـن خـلال تتبـع وإحصـاء خـوادم الويـب النشـطة Active Web Server. ثـم قـام ماتثـوي بعـد ذلـك بتعديـل الربـوت حتـى يتمكـن مـن تجميـع محـددات المصـادر الموحـدة URL`s. وقـد عُـرفـت قاعـدة البيانـات التـي تـم تجميعهـا مـن خـلال هـذا الروبـوت بـ Wandex. وفـي أكتوبـر عـام 1993 قـام أرتيجـن كوسـتر Artijn Koster بتطويـر محـرك جديـد يشـبه فـي بنيتـه المحـرك Archie وعُـرف هـذا المحـرك بـ Aliweb. وقـد أتـاح هـذا المحـرك لأول مـرة إمكانيـة تسـجيل الصفحـات فـي محـركات البحث، حيـث أتـاح الفرصـة لمعـدي صفحـات الويـب أن يقومـوا بتسـجيل الصفحـات وتكشـيفها ووصفهـا بأنفسـهم، ولكنـه واجـه مشـكلة كبيـرة هـي أن معـدي صفحـات ومواقـع الويـب لـم يكـن لديهـم الخبـرة الكافيـة لتكشـيف وتسـجيل صفحاتهـم بأنفسـهم (SEO, 2003).

وبحلـول ديسـمبر عـام 1993 ظهـرت ثلاثـة محـركات بحـث جديـدة فـي الوقـت نفسـه هـي علـى التوالـي: The World Wide Web Worm- WWWW, JumpStation, The Repository-Based Software Engine- RBSE. وقـد اعتمـد المحـرك JumpStation علـى تكشـيف عناويـن ورؤوس الصفحـات Title and Header كمـا اعتمـد فـي الاسـترجاع علـى البحـث الخطـي[2] Linear Search. ومـع نمـو الشـبكة العنكبوتيـة لـم يعـد هـذا المحـرك قـادراً علـى متابعـة هـذا النمـو السـريع ممـا جعلـه يتوقـف سـريعاً. أمـا المحـرك WWW Worm فقـد اعتمـد علـى تكشـيف العناويـن ومحـددات المصـادر الموحـدة Page Title and URL`s. ومـن العيـوب الأسـاسـية فـي كل مـن JumpStation and WWWW أنهمـا كانـا يسـترجعان النتائـج دون أي ترتيـب، حيـث كان يتـم اسـترجاع النتائـج وفقـاً للترتيـب الـذي وجـدت عليـه فـي قاعـدة البيانـات. أمـا المحـرك RBSE فقـد كان أول

(1) **كمبيوتر روبوت**: هو ببساطة برنامج حاسب آلي يستطيع أداء العديد من المهام التكرارية بسرعة كبيرة جداً تفوق إمكانيات مئات بل آلاف الأشخاص إذا حاولوا القيام بالوظيفة نفسها يدوياً.

(1) **البحث الخطي**: هو مضاهاة حروف كلمات الاستفسار حرف بحرف بمعنى أنه إذا كان أحد الحروف غير متشابهة فلا يسترجع أي نتائج وهو يشبه في ذلك البحث بإستخدام CLT + F في الويندوز.

محرك بحث على الشبكة العنكبوتية يستخدم فكرة نظم الترتيب والفرز Ranking Systems والتي يمكن من خلالها استرجاع النتائج مرتبة وفقاً لمعايير الصلاحية[1].

ومع نهاية عام 1993 ظهر المحرك Excite والذي كان ناتج أحد مشروعات تطوير المحرك Architext والذي بدأه 6 طلاب في جامعة ستانفورد في فبراير عام 1993. حيث قاموا باستخدام فكرة التحليل الإحصائي Statistical Analysis لعلاقات الكلمات والمصطلحات Word Relationships من أجل جعل البحث أكثر فعالية وكفاءة (Wall, 2005).

- **الجيل الثاني**

لم تكن كل المحاولات السابقة، في الحقيقة، تمثل مقومات محركات البحث ولم تكن صالحة في الأصل كمحركات؛ نظراً لأن الزاحف Spider أو الروبوت Robot الذي يتولى تجميع الصفحات من الشبكة العنكبوتية لم يكن بالذكاء الكافي الذي يتمكن خلاله من فهم العلاقات القائمة بين الروابط الفائقة Hyperlinks، ومن ثم فإن المستفيد إذا لم يكن يعلم على وجه الدقة عنوان الصفحة التي يرغب في الوصول إليها فإنه كان من الصعب وربما كان من المستحيل عليه الوصول إلى تلك الصفحة.

وفي يناير عام 1994 ظهر أول دليل بحث على الشبكة العنكبوتية الذي عُرف EINet Galaxy. وقد ساعد على نجاح هذا الدليل اشتماله على ملامح البحث التي وفرها كل من جوفر Gopher وتلنت Telnet (وهما معاً كانا يمثلان أهم أدوات الإنترنت في ذلك الوقت)، هذا إلى جانب ملامح البحث في الشبكة العنكبوتية. وقد شهد أبريل عام 1994 مولد دليل البحث Yahoo على يد كل من ديفيد فيلو David Filo وجيري يانج Jerry Yang، والذي لم يكن في بدايته سوى مجموعة من الصفحات والمواقع المخزنة على الحاسبات الشخصية لدى كل منهما.

---

(1) **معايير الصلاحية**: هي معادلات وخوارزميات رياضية تستخدمها محركات البحث لترتيب النتائج وفقاً لعلاقتها بمصطلحات الاستفسار الذي يدخله المستفيد للبحث في الشبكة العنكبوتية.

- **الجيل الثالث**

شهدت الفترة من عام 1994 حتى نهاية العقد الأخير من القرن العشرين ظهور عدد كبير من محركات وأدلة البحث التي تميزت بقدرتها الفائقة على بحث واسترجاع الصفحات والمواقع على الشبكة العنكبوتية كان أبرزها المحركات الثلاثة Google, AltaVista, Alltheweb، وغيرهم. وقد شهدت الفترة من عام 1994 إلى عام 2000 منافسة شرسة بين مجموعة من محركات البحث العالمية على تغطية أكبر قدر ممكن من صفحات ومواقع الويب، حيث شهدت تلك الفترة العديد من دراسات المقارنة بين مدى تغطية محركات البحث لصفحات ومواقع الويب.

وقد شهدت الفترة من عام 2001 إلى 2010 طفرة جديدة في محركات البحث تمثلت في محاولة معظم المحركات الشهيرة في التحول من مجرد محركات بحث إلى بوابات للويب Web Portals. ويشير مصطلح البوابات إلى مجموعة الأدوات التي تسعى إلى تنظيم مصادر المعلومات المتاحة من خلال تقسيمات موضوعية شاملة بحيث تشتمل البوابة على جميع أنواع المصادر والخدمات التي يحتاج إليها المستفيدون من خدمات الشبكة العنكبوتية مثل خدمات البريد الإلكتروني، والدردشة، وقوائم الخدمات والقوائم البريدية، والمواد الإخبارية، وأسعار العملات، وأحوال الطقس، إلى جانب قوائم موضوعية بمصادر المعلومات المتاحة من خلال البوابة إلى جانب محرك يتيح إمكانية البحث في البوابة. وإلى جانب التنوع في الخدمات التي تقدمها البوابات للمستفيدين منها نجد أن هذه المواقع عادة ما تتضمن برامج تساعد على تحليل استخدامات المستفيدين Web Usage Analyzer بغرض بناء ملفات سمات المستفيدين User Profiles ويمكن من خلال هذه الملفات التعرف إلى احتياجات المستفيدين والتنبؤ بها، بالتالي اختيار المصادر المناسبة لكل مستفيد من المستفيدين من الموقع. ويمكن أن تقوم تلك المواقع باستخدام تكنولوجيا الدفع Pushing Technology إلى المستفيدين من الموقع. كما يمكن أن تتم عملية الدفع عبر خدمات البريد الإلكتروني التي توفرها تلك المواقع أو إلى الصفحات الأمامية للمستفيدين من هذه المواقع كما يمكن أن يتم الدفع إلى دوسيهات خاصة للمستفيدين

من هذه المواقع. من ثم فالبوابات عادة ما تيسر لمستخدمي تلك المواقع كل أنواع الخدمات التي يحتاجون إليها بصورة تفاعلية، مما يوفر كل احتياجات المستفيد من خدمات ومصادر الشبكة العنكبوتية. وفي مقابل ذلك تسعى البوابات إلى جذب الشركات التي تسعى إلى الإعلان عن منتجاتها وخدماتها لتحقيق الأرباح من خلال تلك المواقع، حيث إنه من المعروف أنه كلما زاد عدد مستخدمي الموقع، تهافتت الشركات على الإعلان عن خدماتها ومنتجاتها من خلال هذه المواقع.

- **الجيل الرابع**

شهدت الفترة من عام 2000 بداية تطوير جيل جديد من أدوات البحث على الشبكة العنكبوتية يعرف بالأعوان الذكية للبحث Intelligent Agent التي تسعى إلى الاستفادة من إمكانيات الذكاء الاصطناعي والنظم الخبيرة لتحقيق متطلبات تشغيل الويب الدلالي Semantic Web في تيسير عمليات البحث والاسترجاع ومازال العمل في هذه الأدوات في طور التجارب المبدئية.

وتتنوع طرق الوصول إلى مصادر المعلومات المتاحة على الشبكة العنكبوتية بين أربعة أساليب أساسية هي (Vaughan, & Thelwall, 2003; Gordon & Pathak 1999).

10.1 الإبحار Navigation

يستخدم الإبحار آليات الوصول المباشر من خلال أدوات التصفح المعروفة مثل

Internet Explorer أو Google Chrome وما توفره من إمكانيات مثل الإبحار من خلال سطر معين المصادر الموحد URL Line أو الاعتماد على تخزين المواقع المفضلة في ملف المواد المفضلة أو في ملف تاريخ الاستخدام Bookmarks أو Navigation History.

10.2 التصفح Browsing

تنبع تلك الطريقة من طبيعة صفحات الويب التي تقود إلى بعضها البعض من خلال سلسلة متشابكة من الروابط الفائقة. وقد تم توظيف هذه السمة التي تتميز بها

الشبكة العنكبوتية في بناء فهارس موضوعية مصنفة لصفحات الويب تعرف بالأدلة. وهي عبارة عن قوائم برؤوس موضوعات عريضة وتحت كل رأس موضوعي عريض مجموعة من الرؤوس الثانوية التي تقود إلى صفحات الويب المرتبطة بالرأس الثانوي مرتبة وفقاً لقوة العلاقة بين الصفحة والرأس. بالطبع يمكن لهذه الأدلة أن تقوم بتكشيف الصفحة نفسها تحت أكثر من رأس موضوع واحد.

10.3 أدوات البحث والاسترجاع على الويب Web Searching and Retrieval Tools

وتنقسم تلك الأدوات إلى ثلاثة أنواع رئيسة هي:

10.3.1 أدلة البحث

في عام 1994 قام ثنان من طلبة الدكتوراة بجامعة هارفرد هما جيري يانج وديفيد فيلو Yang and David Filo يدوياً بتنظيم مجموعة من صفحات الويب التي كانت متاحة على حواسيبهم الشخصية في شكل دليل. وقد تطور هذا الدليل سريعاً ليصبح أشهر دليل بحث على الويب وقد أطلقا عليه دليل البحث ياهو Yahoo. ويتيح دليل البحث إمكانية الإبحار وتصفح مواقع الويب بالاعتماد على بنية هرمية مصنفة للويب (Gulli & Signori, (2005. فعلى سبيل المثال عند البحث عن موقع عن تاريخ الويب يجب على الباحث التزام التتابع التالي لكي يصل إلى المعلومة المطلوبة:

Computer and Internet > Internet > World Wide Web > History

وعلى الرغم من أن عملية البحث من خلال التزام بنية هرمية ثابتة تساعد على الوصول إلى المعلومات المطلوبة أحياناً خاصة عندما يكون الباحث على دراية بالموضوعات وعلاقاتها بعضها بعضاً، إلا أنها لا تصلح لتلبية كل الاحتياجات البحثية فنفترض مثلاً أن أحد الباحثين يريد معلومات عن «من هم مؤسسو دليل البحث ياهو؟» في هذه الحالة فإن عملية الوصول للمعاومات المطلوبة قد تستغرق وقتاً طويلاً نظراً لأن الباحث بحاجة إلى البحث في البنية الهرمية للدليل ثم تصفح كل

الصفحات المسترجعة للوصول إلى المعلومة المطلوبة. هذا إضافة إلى أن عملية بناء أدلة البحث تعتمد على تجميع صفحات الويب يدوياً وتكشيفها يدوياً، مما يتعذر معه تغطية كل الصفحات، كما أنه يحتاج إلى وقت طويل للتعرف إلى الصفحات الجديدة والتعديلات التي تجرى على الصفحات القديمة. من هنا ظهرت الحاجة إلى أدوات أكثر سرعة في تغطية النمو الهائل في صفحات الويب، إضافة إلى متابعة التغييرات التي تجرى على هذه الصفحات. وقد كان لظهور وتطور محركات البحث أكبر الأثر في حل تلك المشكلة.

10.3.2 محركات البحث
Search Engines

تعمل محركات البحث بصفة أساسية على بناء كشافات لمصادر المعلومات المتشابكة من خلال اشتقاق كلمات أو عبارات من النصوص نفسها لبناء ملفات تسمح ببحث هذه المشتقات بالاعتماد على أساليب البحث والاسترجاع المعروفة مثل المنطق البوليني، وتجاور المصطلحات، والبتر، والجذع وغيرها. والحقيقة أن هذه الملفات لا تتميز عن الأساليب التقليدية التي استخدمت في الاسترجاع منذ أن حل الاسترجاع العشوائي محل الاسترجاع التسلسلي، والتي تشتمل بصفة أساسية على ثلاثة ملفات حيوية هي: الملف التسلسلي Serial File، والملف الكشفي Index File، والملف المقلوب Inverted File. ومع ذلك فإن التقنيات الحديثة من أجهزة وبرمجيات ساعدت على تحديث وبحث تلك الملفات المقلوبة بسرعة كبيرة، هذا إلى جانب أنها أضافت إلى تلك الملفات مجموعة جديدة من الملفات لتيسير عمليات البحث والاسترجاع مثل ملف الروابط الفائقة، ملف وصف الوثائق.. إلخ (Lancaster, 1998).

- **الفرق بين محركات وأدلة البحث**

قبل التعرف إلى طريقة عمل محركات البحث لا بد من التمييز بين محركات وأدلة البحث وما هي المتطلبات التي دفعت إلى التنويع في أدوات البحث والاسترجاع.

الملمح الأساسي الذي يميز محركات البحث عن أدلة البحث أنها تعتمد بشكل أساسي على برامج الزحف Crawling Software التي تقوم بمسح الشبكة العنكبوتية للتعرف إلى الصفحات الجديدة وتجميع نسخ منها في ملفات خاصة من أجل تيسير عمليات تكشيفها. هذه الزواحف عبارة عن برامج تقوم بتتبع الروابط الفائقة من صفحة إلى أخرى ومن موقع إلى آخر. وفي بعض الحالات يمكن لصاحب الموقع أن يُعرف محرك البحث على موقعه من خلال تعريف العنكبوت أو الزاحف على عنوان هذا الموقع أو معين المصادر الموحد (URL) الخاص بهذا الموقع. أما الأدلة فهي لا تعتمد على برامج للزحف، وإنما تعتمد بشكل أساسي على الإمكانيات البشرية في تصفح الشبكة العنكبوتية للتعرف إلى الصفحات الجديدة وتكشيفها.

لذلك يمكن القول إن محركات البحث تعتمد على التجميع والتكشيف الآلي، بينما تعتمد أدلة البحث على التجميع والتكشيف اليدوي. بالتالي فإن محركات البحث تستطيع التجميع والتكشيف بسرعة أكبر بكثير من سرعة أدلة البحث مما يجعلها أكثر شمولاً في تغطية صفحات ومواقع الويب.

ويتبادر إلى الذهن هنا سؤال مهم هو لماذا نحتاج إلى أدلة بحث ما دامت محركات البحث أكثر سرعة وكفاءة؟

الإجابة بشكل مختصر هي الجودة Quality حيث إن القائمين على تجميع الصفحات وتكشيفها بشكل يدوي بالطبع لديهم قدرة أكبر على التمييز بين الصفحات والتعرف إلى مدى ملاءمتها للفئة التي يتم تصنيف الصفحة تحتها. كما أن هذا الشخص لديه قدرة أكبر من البرامج على تجميع الصفحات المهمة واستبعاد الصفحات غير المهمة واختيار الرؤوس المناسبة. وقد أثبتت التجارب العلمية العديدة التي أجريت للمقارنة بين أساليب التكشيف اليدوي والتكشيف الآلي تفوق التكشيف اليدوي في دقة النتائج المسترجعة عن التكشيف الآلي، بينما يتفوق التكشيف الآلي في عدد النتائج المسترجعة.

في عام 2008 سجل محرك البحث جوجل أنه اكتشف أكثر من تريليون معين مصادر موحد Uniform Resources Locators - URLs لصفحات ومواقع ويب قابلة

للبحث والاسترجاع من خلال المحرك. ومع ذلك أشار العديد من الدراسات إلى أنه لا يوجد محرك بحث واحد قادر على تكشيف وبحث كل صفحات الويب المتاحة على الإنترنت. وسنعرض فيما يلي كيف تعمل محركات البحث على تيسير بحث واسترجاع صفحات الويب من خلال عرض عمليات التجميع والتكشيف والعوامل التي تؤثر في البحث وترتيب الصفحات المسترجعة (1.2). حيث تعتمد محركات البحث على تجميع صفحات الويب من خلال أدوات يطلق عليها الزواحف التي تقوم بالحصول على نسخ من صفحات الويب ثم تقوم المحركات بتكشيف تلك الصفحات وإعداد كشافات تيسر عمليات البحث والاسترجاع من خلال أدوات البحث التي يستخدمها الباحثون أثناء التفاعل مع واجهات تعامل متاحة من خلال الويب. من ثم فمحركات البحث تتكون من 5 عناصر أساسية هي: الزواحف، والكشافات، وقاعدة البيانات، وأداة البحث، وواجهة التعامل إلى جانب آليات الفرز والترتيب.

I. زواحف الويب Web Crawling:

تعد أداة ماثيو جاري Matthew Gray التي طورها خلال عام 1993، والمعروفة بـ World Wide Web Wanderer، أول محاولة لتطوير أداة للتجميع الآلي لصفحات الويب في مقابل التجميع اليدوي الذي اعتمدت عليه أدلة البحث (Gray, 1995). واعتمدت تلك الأداة على تحميل صفحات الويب واختبار الروابط الفائقة التي تربطها بصفحات أخرى ثم تقوم بتحميل كل الصفحات المرتبطة التي تكتشفها أثناء تتبع روابط الصفحة الأصلية حتى تنتهي من تجميع كل الصفحات التي تكتشفها أثناء عملية التصفح. وهي الطريقة التي تعمل بها كل أدوات التجميع الآلي والتي يطلق عليها العنكبوت Spider أو الروبوت Robot.

ونظراً لضخامة حجم الويب فإن محركات البحث عادة ما توظف آلاف الزواحف التي تقوم بتصفح الشبكة العنكبوتية لتحميل صفحات الويب، والبحث عن روابط فائقة لصفحات جديدة، إضافة إلى إعادة زيارة الصفحات القديمة التي يمكن أن يكون محتواها قد تغير. وعادة ما تعتمد محركات البحث على زيارة الصفحات بناء على معدلات وتتابع التغيير في تلك الصفحات وذلك بغرض تحديث محتوى الكشافات التي تتضمن معلومات عن تلك الصفحات.

وتعدُّ تغطية كل ما تحويه الويب من صفحات أمراً في غاية الصعوبة ومن التحديات التي لم تستطع أي أداة إلى اليوم التغلب عليها، ليس فقط بسبب حجم الويب ولكن أيضاً بسبب معدلات التغيير السريعة في محتوى صفحات ومواقع الويب. كما أن العديد من الصفحات تظهر وتختفي بمعدلات سريعة، وهو ما يطلق عليه الروابط الفائقة غير النشطة Inactive Link Died Link،. ويرى بروستيلر كاهلي Brewster Kahle مؤسس أرشيف الإنترنت Internet Archive أن العمر المتوقع لأي صفحة ويب قد يصل إلى 100 يوم في المتوسط (Weiss, 2003).

وتنقسم الويب إلى ثلاثة مستويات من حيث إمكانيات تعامل الزواحف مع تلك الأدوات (Bergman, 2001):

- **الويب السطحي Surface Web:**

ويطلق عليه أيضاً مستوى الويب المرئي Visisble Web أو الويب المكشف Indexable Web أو الويب المضيء Lighened Web ويشمل جزءاً من الشبكة العنكبوتية العالمية المتاحة للمستفيد العام دون الحاجة إلى تحقق من هوية المستفيد كما أنه متاح للتجميع من خلال الزواحف والتكشيف بمحركات البحث.

- **الويب العميق Deep Web:**

يطلق عليه مستوى الويب غير المرئي أو الويب المخفيInvisisble Web وهو أجزاء من شبكة الويب التي لا تتمكن زواحف الويب من الوصول إليها وتكشيف محتوياتها بمحركات البحث. وعادة ما تستخدم المواقع الحكومية والتجارة الإلكترونية ومواقع البنوك والجامعات هذا الجزء من الويب.

- **الويب المظلم Dark Web:**

شبكة الويب المظلمة هي جزء مخفي من الإنترنت لا يمكن الوصول إليه إلا باستخدام برامج خاصة مثل TOR OR The Onion Router، وهي شبكة تصفح شُعَبِيَّة مجهولة تستخدم للاتصال بالويب المظلم. وعادة ما يستخدم قراصنة الويب

هـذا الجـزء مـن الويـب لإخفـاء برامـج التجسـس التـي يسـتخدمونها حتـى لا تتمكـن أدوات البحث من اكتشـافها واكتشـاف مصدرهـا كمـا تسـتخدمها المواقع غيـر القانونيـة فـي بـث معلوماتهـا.

ويمكن تقسيم زواحف الويب إلى ثلاثة أنواع هي:

أ. الزواحف الآلية Automated Based Crawlers

هـي الزواحـف التي تعتمـد عليهـا محـركات البحـث فـي اصطيـاد الصفحـات وتجميعها بصـورة آليـة دون تدخـل بشـري. وتسـتخدم تلـك الزواحـف برامـج حاسـب آلـي تقـوم بتصفـح الويـب لتحديـد الصفحـات الجديـدة ثـم تقـوم باصطيادهـا وتجميعهـا.



شكل (1/10) نموذج مبسط للويب بمستوياتها الثلاثة

Deep Web Technology. https://www.deepwebtech.com/deepweb-not-darkweb

ب. الزواحف البشرية Human Based Crawler

وتعتمد على آليات التجميع اليدوي التي توظفها أدلة البحث من خلال مجموعة من مجمعي الصفحات الذين يجوبون الشبكة العنكبوتية لاصطياد الصفحات وتجهيزها للتكشيف والبحث فيها.

ت. الزواحف المختلطة Hybrid Crawlers" Or Mixed Results

تعتمد عليها بعض محركات البحث للتأكد من تجميع صفحات الويب السطحي والعميق معاً إلا أنها نادرة الاستخدام لارتفاع تكلفتها ومن أمثلة محركات البحث التي تعتمد على هذا الأسلوب أداة Inktomi.

وتحدد بعض الصفحات التي لا يرغب القائمون عليها إتاحتها من خلال محركات البحث وذلك لأسباب متعددة منها: أن تشتمل على معلومات خاصة بالعاملين في مؤسساتهم فقط، أو تتضمن معلومات لها درجة سرية محدودة أو غيرها من الأسباب. وفي هذه الحالة يستبعد القائمون على تطوير هذه الصفحات تجميعها من خلال الزواحف باستخدام بروتوكول استبعاد الروبوت Robots Exclusion Protocol وهو عبارة عن كود يتم وضعه ضمن أكواد HTML بالصفحة لاستبعاد الزواحف من التعامل مع تلك الصفحة.

أما الغالبية العظمى من المؤسسات فترغب في تكشيف وإتاحة صفحاتها من خلال محركات البحث، ما يعطيها فرصة أكبر للظهور والاسترجاع. فيقوم المسؤولون عن تطوير الصفحة باستخدام بروتوكول خريطة الموقع Sitemap Protocol وهو أداة تدعمها معظم محركات البحث تتيح للزاوحف قائمة بعناوين المصادر الموحدة التي يمكن تكشيفها عند التعامل مع الموقع (https://www.sitemaps.org/). وتعد هذه التقنية في غاية الأهمية للزاوحف حيث تمكنها من التعرف إلى عناوين المواقع التي لا يمكنها الوصول إليها من خلال أساليب الزحف التقليدية بالتالي لا يمكنها الوصول إلى تلك الصفحات وخاصة صفحات الويب العميق.

ويعرض الجزء التالي كيف تقوم محركات البحث بتكشيف الصفحات وتيسير استرجاعها عندما يقوم الباحث بإدخال مصطلحات الاستفسار في واجهة البحث.

II. التكشيف والفرز Indexing and Ranking

عندما ينتهي الزاحف من اصطياد الصفحات ويقوم بتجميعها في مستودع الوثائق يقوم محرك البحث بتكشيف محتوى الصفحات، حيث يقوم بتجميع الكلمات والمصطلحات والعبارات الواردة في تلك الصفحات مع استبعاد الكلمات كثيرة التردد والتي يطلق عليها كلمات الوقف Stop Words وهي الكلمات التي تتردد كثيراً في الوثائق لتكملة السياقات. وعادة ما تفتقر هذه الكلمات إلى الدلالة الموضوعية التي يمكن استخدامها في البحث عن الوثيقة مثل حروف الجر وأسماء الإشارة والمكان والزمان سواء كان ذلك للوثائق باللغة العربية (في، من، على، عند.. ..الخ) أو باللغة الإنجليزية (a, an, the, when, on… etc). كما تقوم المحركات أيضاً باستخدام أسلوب الجذع Stemming وهو عبارة عن طريقة تساعد على تجميع الأصول اللغوية للكلمات والمصطلحات من خلال استبعاد البدايات Prefixes واللواحق Suffixes مما يساعد على تحسين مستوى تكشيف الكلمات وبناء كشافات أكثر دلالة على المحتوى الموضوعي للوثائق. فعلى سبيل المثال كلمات مثل eating, eats and eaten كلها مشتقات من الأصل اللغوي eat بالتالي فإن البحث عن المصطلح eat سوف يسترجع كل المشتقات وبدائل والمصطلح مما يحسن من كفاءة الكشاف.

ويمكن تصور شكل الكشاف بأنه عبارة عن قائمة بالمصطلحات الواردة في صفحات الويب وأمام كل مصطلح من هذه المصطلحات أرقام الوثائق التي ورد بها المصطلح القابل للبحث. فعلى سبيل المثال إذا كان الكشاف يشتمل على أربعة مصطلحات وأرقام الوثائق التي تعبر عنها هي كالتالي:

جدول رقم (10.2) المصطلحات الكشفية وطريقة تمثيلها بالمحركات

المصطلحات بالكشاف	أرقام الوثائق
Internet	2,5
Search	1,5,6
Browse	1,2
Tool	4

فإذا كان الباحث يبحث عن المصطلح Search فإن النتائج المسترجعة ستشتمل على الوثائق 1, 5 ,6 بينما البحث عن Internet Search فسينتج عنه استرجاع الوثيقة رقم 5 فقط، حيث إنها الوثيقة الوحيدة التي ورد بها كل من المصطلحين معاً. وذلك في حالة اعتبار المعامل AND هو المعامل الرئيس عند البحث بجمل. وقد تعتمد محركات البحث على معاملات أخرى، وسوف يتم مناقشة أساليب البحث بشكل أكثر تفصيلاً فيما يلي.

وتستخدم محركات البحث أيضاً أساليب لوزن المصطلحات عند بناء الكشافات تعتمد عليها في ترتيب الوثائق المسترجعة. ويوجد أكثر من طريقة تعتمد عليها محركات البحث لوزن الوثائق والمصطلحات نذكر منها ما يلي:

الوزن Weighting: يقوم على تحديد قيمة رقمية للمصطلح تحدد مدى صلاحيته وأهميته بالنسبة للوثيقة التي تم تكشيف المصطلح منها. ومن أبرز أساليب وزن المصطلحات استخدام عدد مرات تردد المصطلح في الوثيقة Term Frequency والذي يتم على أساسه تحديد أهمية المصطلح بالنسبة للصفحة وفقاً لعدد مرات تردد المصطلح في الصفحة. فعلى سبيل المثال إذا كان أحد الباحثين يريد معلومات عن Egypt فإن الصفحة التي يرد فيها المصطلح Egypt خمس مرات عادة ما تكون أكثر أهمية من صفحة أخرى يرد فيها المصطلح مرة واحدة. وعلى الرغم من ذلك فإن تردد المصطلحات يتأثر بعاملين أساسيين هما(Garcia-Molina & Gyngyi, 2004) .

حجم الصفحة Page Size

فعلى سبيل المثال الصفحة التي تردد المصطلح بها 5 مرات، وتشتمل على 1000 كلمة تصبح أهمية المصطلح بالنسبة لهذه الصفحة تعادل 0.005 . بينما الصفحة التي تردد بها المصطلح مرة واحدة وتشتمل على 100 كلمة فقط، تكون أهمية هذا المصطلح بالنسبة لهذه الصفحة هي 1 ٪، من ثم تكون الصفحة التي ورد بها المصطلح مرة واحدة أكثر أهمية من صفحة أخرى ورد بها المصطلح 5 مرات نظراً لأن حجم الصفحة أثر في الأهمية النسبية للمصطلح.

1. الخداع Spamming

استخدام تردد المصلحات كأسلوب لتحديد الأهمية النسبية لصفحات الويب يتأثر بأساليب إغراق الصفحات بكلمات ومصطلحات وتكرارها عدد من المرات لزيادة الأهمية النسبية لهذه الصفحات عند مقارنتها بصفحات أخرى. فعلى سبيل المثال إذا أراد مطورو صفحات الويب أن يتم تكشيف الصفحة التي يقومون بإعدادها تحت مصطلح أو مجموعة معينة من المصطلحات، فإنهم يكررون هذا المصطلح عدداً كبيراً من المرات لزيادة الأهمية النسبية للوثيقة عند تكشيفها تحت هذا المصطلح، مما يرفع من مكانتها في الترتيب النهائي للوثائق. ويعرف هذا الأسلوب بخداع محركات البحث Search Engine Persuasion.

2. الترتيب وفقاً لموقع المصطلح وشكله

هذه الطريقة تعتمد على إعطاء وزن نسبي للصفحة بناء على السياق الذي ورد فيه المصطلح في الصفحة، فإذا ظهر المصطلح في الصفحة مكتوباً بخط كبير أكبر أو أعرض Large or Bold من بقية المصطلحات فإن ذلك يعني أن هذا المصطلح له أهمية نسبية أكبر من غيره من المصطلحات. كما أن ظهور المصطلح في أماكن معينة مثل عنوان الوثيقة قد يعني أن المصطلح له قيمة أكبر من غيره من المصطلحات التي لم ترد بعنوان الوثيقة.

3. استخدام نصوص الزاوية Anchor Text

تعتمد هذه الطريقة على إعطاء أهمية نسبية للوثيقة وفقاً لعدد مرات ظهور المصطلح ضمن أقواس الزاوية للوثيقة المصدرية أو ضمن أقواس الزاوية لوثيقة أخرى تشير إلى الوثيقة. بعبارة أخرى إذا كان المصطلح ورد بالوثيقة وبه رابطة نشطة لصفحة أخرى فإن ذلك يعني أنه مصطلح مهم، كما أن ورود المصطلح بوثيقة أخرى بها رابطة نشطة تشير إلى الوثيقة المكشفة يعني أن الوثيقة الحالية تتناول المصطلح المشار إليه من وثيقة أخرى.

على سبيل المثال إذا كانت الوثيقة الحالية بها رابطة نشطة لمصطلح Search

Engines فإن هذا يحمل معنيين: أن هذا المصطلح مهم بالنسبة للوثيقة الحالية كما أنه أيضاً مهم بالنسبة للوثيقة التي يشير إليها.

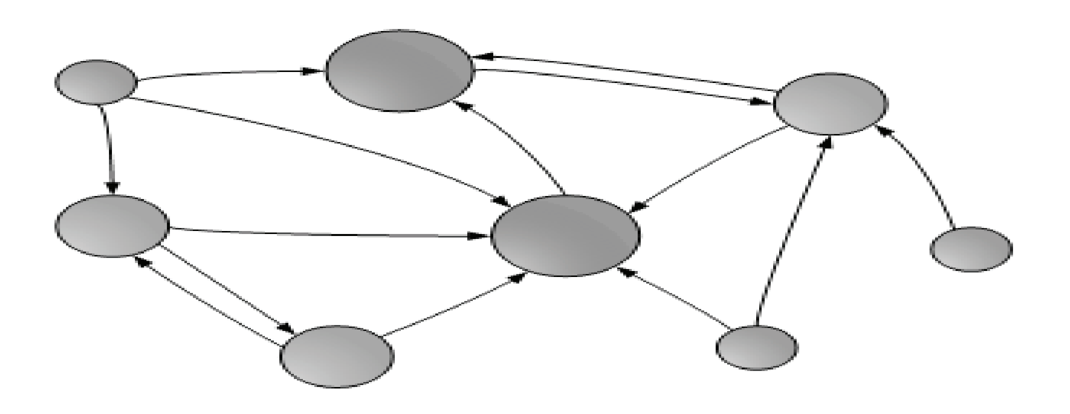
وقد أدى استخدام محركات البحث لهذا الأسلوب إلى ظهور ما يعرف بالروابط المخادعة Spamming Links وخاصة لدى محرك البحث جوجل فيما عرف بفرقعات جوجل Google Bombing، ولعل أشهر أمثلة فرقعات جوجل التي جاءت نتيجة لاستخدام أسلوب تحليل نصوص الزاوية لاسترجاع صفحة البيت الأبيض White House في قمة النتائج المسترجعة عند البحث في جوجل عن مصطلح miserable failure وهي الفضيحة التي اهتم بها الإعلام الأمريكي، نظراً لوجود الكثير من صفحات الويب التي تشير إلى موقع البيت الأبيض باستخدام هذا المصطلح ضمن نصوص الزاوية الخاصة بها. وقد عالج جوجل خلال السنوات القليلة الماضية مشكلة الفرقعات من خلال تطوير خوارزميات التكشيف وآليات البحث (Moulton & Carattini, 2009).

4. استخدام الروابط الفائقة

يعتبر استخدام الروابط الفائقة لرسم شكل الويب من أكثر الأساليب استخداماً بمحركات بحث الشبكة العنكبوتية، حيث يعتمد هذا الأسلوب على عرض الشبكة العنكبوتية في صورة نقاط ارتكازية يطلق عليها أسانيد Authorities وروابط Links توضح صورة بيانية لصفحات الويب وعلاقتها ببعضها بعضاً. فقد قام كل من سيرجي براين ولاري بيدج Sergey Brin and Larry Pag، عندما كانا طلبة دكتوراة بجامعة ستانفورد بتطوير محرك البحث جوجل، بالاعتماد على فكرة رسم الويب في صورة شكل من خلال توضيح علاقة صفحات الويب ببعضها البعض مما يساعد في تحديد صلاحية صفحات الويب من خلال دراسة تلك العلاقات. ففي عام 1998 قاما بإعداد دراسة عن كيفية قياس صلاحية صفحات الويب من خلال دراسة موقع صفحة الويب في إطار الشكل العام للويب Web Graph وبصفة خاصة عدد الروابط الفائقة المرتبطة بالصفحة Incoming Links وعدد الروابط الفائقة الخارجة من الصفحة Outgoing Links. وتعتمد هذه الطريقة على فكرة الاستشهادات المرجعية التي استخدمها يوجين جارفيلد Eugene Garfield في تحديد الأهمية النسبية

للدوريات العلمية والأهمية النسبية للمقالات ومؤلفي المقالات، حيث يتم تقييم الصفحة على أساس عدد الاستشهادات (الروابط التي تشير منها وإليها). فالصفحة التي تتلقى عدداً كبيراً من الاستشهادات في موضوع معين تعد صفحة أكثر أهمية من صفحة أخرى تتلقى عدداً أقل من الاستشهادات، بالتالي فالصفحة التي تتلقى عدداً كبيراً من الاستشهادات لا بد أن يتم ترتيبها أعلى من الصفحة التي تتلقى عدداً أقل من الاستشهادات. وقد أطلق براين وبيدج على خوارزمية الفرز مصلح ترتيب الصفحة PageRank والتي تمثل الأداة الأساسية في بنية محرك البحث جوجل (Brin & Page, 1998). وقد بدأ معظم محركات البحث منذ بداية الألفية الجديدة الاعتماد على تحديد الرسم البياني للويب كأداة أساسية في إعداد خوارزميات الترتيب التي تستخدمها في ترتيب النتائج.

ويوضح الشكل رقم (10.3) الرسم البياني للويب حيث تظهر فيه مجموعة من الصفحات على أنها نقاط ارتكازية والروابط المرتبطة بهذه الصفحات. ويتم تحديد ترتيب الصفحة بناء على حجم ولون النقاط الارتكازية، ومن الملاحظ أن الصفحات التي حصلت على ترتيب عالٍ High PageRank (والممثلة باللون الأحمر) هي الصفحات التي تشتمل على عدد أكبر من الروابط عن الصفحات ذات الترتيب المنخفض Low PageRank والممثلة باللون الأخضر.



شكل رقم (10.3) رسم بياني مبسط للويب يوضح طريقة تحديد ترتيب الصفحة PageRank

III. قواعد البيانات Databases

تعد قواعد البيانات، التي يُطلق عليها أحياناً مستودعات الوثائق، المصدر الأساسي للمعلومات أثناء عمليات البحث والاسترجاع في تلك المحركات، ومع ذلك فهي لا تمثل بديلاً للشبكة العنكبوتية، وإنما تتضمن معلومات عن الصفحات، هذه المعلومات تساعد محركات البحث على إجراء عمليات البحث والاسترجاع، وعادة ما يطلق على هذه المعلومات النقاط الكشفية Indexing Points. ومع ذلك فهناك مجموعة من محركات البحث التي تحتفظ بنسخ كاملة من صفحات الويب التي تقوم بتكشيفها مثل محرك البحث جوجل Google ومحرك البحث Alltheweb حيث يقوم كل منهما ببناء مستودعات كاملة بكل الصفحات التي يتم تكشيفها لتيسير عمليات المتابعة والتحديث. كما أن هذه المستودعات تفيد كثيراً في حالة حذف الصفحة من الخادم الرئيس، حيث يمكن استرجاع الصفحة من أرشيف محرك البحث من خلال ما يعرف بالصفحة المخبأة Page Cash. ويعمل محرك البحث جوجل الآن على بناء أرشيف للإنترنت بالصفحات التي تتضمنها قاعدة بياناته، ويتم تحميل هذا الأرشيف بالعديد من المؤسسات للحفاظ على تاريخ الإنترنت.

IV. برامج البحث Search Software

تعد برامج البحث والاسترجاع من أكثر المكونات أهمية بالنسبة لمستخدمي محركات البحث، حيث إن هذه البرامج هي التي تقرر أي الصفحات تتناسب مع استراتيجية البحث أو السؤال الذي يوجه المستفيد لمحرك البحث، كما أنها أيضاً تحدد ترتيب الصفحات المسترجعة، حيث تدفع هذه البرامج بالصفحات الأكثر أهمية إلى قمة القائمة، تليها الصفحات الأقل أهمية فالأقل. ويتم ذلك بناء على مجموعة من المعادلات الرياضية التي تعرف في مجال استرجاع المعلومات بخوارزميات محركات البحث Search Engines Algorithms.

ويقوم العديد من المتخصصين في عمليات رفع كفاءة محركات البحث (Search Engines Optimization (SEO بقضاء وقت طويل في محاولة منهم لفهم الطرق التي تستخدمها محركات البحث في ترتيب الصفحات المسترجعة من أجل وضع تعليمات

تساعد على رفع ترتيب الصفحات ضمن النتائج المسترجعة. كما تتضمن تلك البرامج الأساليب المختلفة التي يمكن للمستفيد أن يستخدمها في إعداد استراتيجية البحث أو صياغة الاستفسار بطريقة تساعد الباحث على الوصول إلى أفضل النتائج.

وتجدر الإشارة إلى أن محركات البحث عادة ما تعد الأساليب التي تستخدمها في وزن المصطلحات وترتيب الصفحات من الأسرار التي لا يمكن نشرها حيث إنها تعد الميزة التنافسية التي تميزها عن غيرها من محركات البحث، كما أن إعلانها لمطوري مواقع وصفحات الويب قد يؤدي إلى اتباع طرق تؤدي إلى خداع تلك المحركات. ومع ذلك فإن الشركات والمؤسسات التجارية تهتم كثيراً بترتيب مواقعها في محركات البحث فيما يعرف بصفحة نتائج محرك البحث (Search Engine Result Page (SERP، نظراً لأن المستفيدين عادة ما يهتمون فقط بالصفحة الأولى من نتائج البحث ويقومون بعرض عدد محدود جداً من النتائج المسترجعة في قمة هذه الصفحة وتجاهل النتائج التي تظهر في ذيل صفحة نتائج البحث. وتلعب صفحة نتائج البحث في المحركات دوراً أساسياً في دعم أهمية مواقع الشركات والإعلان عنها حيث إنها لها حوافز اقتصادية كلما كان الموقع يظهر ضمن المجموعة الأعلى ترتيباً Highly Ranking ضمن النتائج المسترجعة. لذلك تقوم الشركات بشراء مساحات وأماكن معينة لعرض إعلاناتها في صفحة نتائج محركات البحث فيما يعرف بالنتائج المدعومة (الرعاية) (sponsored results (Cutts, 2006.

وتوجد صناعة قائمة على ما يعرف بالترقية في محركات البحث (Search Engine Optimization (SEO تتيح للعديد من الشركات القيام بمجموعة من الإجراءات التي تساعد على تحسين ترتيب صفحات الويب ضمن صفحة نتائج محركات البحث بالاعتماد على الأساليب التي تمت مناقشتها أعلاه مما يساعد أيضاً على زيادة عدد الروابط الفائقة وجودة تلك الروابط.

وتعرف الترقية بأنها أسلوب أو طريقة يمكن من خلالها لمواقع وصفحات المعلومات المتاحة على الشبكة العنكبوتية أن تحصل على ترتيب (Ranking) أعلى في محركات البحث. ويوجد العديد من المصطلحات التي تستخدم للدلالة على

الترقية في محركات البحث منها ترتيب محركات البحث Search Engine Ranking، والترقية من خلال محركات البحث Search Engine Promotion، وترقية مواقع المعلومات Website Promotion وإزعاج الكشاف Spam Index ومزرعة الروابط Link Farm وذلك من خلال حشو صفحات الويب بعدد كبير من الروابط لخداع محركات البحث. وعندما تكتشف محركات البحث هذا السلوك تقوم بعقاب صفحة الويب من خلال استبعاد الصفحة من الكشاف وحظر حصادها وتجميعها من خلال الزاحف بالتالي تكشيفها لفترة زمنية معينة (Cutts, 2006).

ويشير العديد من دراسات المستفيدين من محركات البحث إلى أن 1 من كل 20 مستفيداً يتعاملون مع النتائج التي تظهر في الصفحة الثانية من نتائج البحث وأن 1 من كل 100 مستفيد يذهب إلى ما وراء الصفحة الثانية. ويوجد العديد من العوامل التي تدفع المؤسسات إلى الترقية في محركات البحث منها ما يلي:

أ. أسباب اقتصادية: حيث إن ظهور موقع المؤسسة ضمن الصفحات العشر الأولى في محركات البحث يعد من أهم أساليب الدعاية عن المنتجات والخدمات التي تقدمها المؤسسات، مما يساعد على تحفيز الموقع الاقتصادي للمؤسسة وزيادة ربحيتها إذا كانت تهدف للربح.

ب. أسباب سياسية، حيث إن ظهور الموقع ضمن قائمة المواقع في الصفحة الأولى لنتائج البحث يؤدي إلى تمييز هوية المؤسسة Organization Identity في البيئة الإلكترونية، والذي قد يعد أحد الأهداف السياسية للدول التي تساعد على السيطرة من قبل مؤسسات تلك الدولة في قطاعات معينة.

ج. أسباب ثقافية وعلمية مثل كثرة الرجوع إلى مقالات جريدة معينة أو صفحات جامعات أو أشخاص معينين، ما يعزز المكانة الثقافية والعلمية لتلك المؤسسات إضافة إلى حرص العديد من المؤسسات على تقديم المعلومات الصحيحة لجمهور الإنترنت حتى لا يتم خداعهم بمعلومات مضللة وغير حقيقية.

V. واجهة التعامل The Interface

واجهة التعامل هي الجزء الذي يراه المستفيد عند التعامل مع محركات البحث والتي تتضمن صندوق البحث الذي يدخل فيه المستفيد سؤاله، إضافة إلى إعلانات محرك البحث. وعادة ما يبدأ البحث من واجهة التعامل حيث يقوم المستفيد بكتابة استفساره في صندوق البحث، الذي يُرسل مباشرة إلى برامج البحث، التي تقوم بدورها بالبحث في قاعدة البيانات لتحديد كل الصفحات الصالحة للإجابة عن استفسار أو سؤال المستفيد، ثم تتولى بعد ذلك فرز هذه النتائج من الأكثر إلى الأقل صلاحية. ويقوم محرك البحث بإرسال بيانات عن تلك النتائج المرتبة إلى المستفيد وذلك من خلال واجهة التعامل التي استخدمها المستفيد في إعداد الاستفسار. وهذه العملية لا تستغرق أكثر من جزء من الثانية مما يوحي بمدى سرعة المحركات في أداء عمليات البحث والاسترجاع، وهو ما يعطيها قيمة وأهمية كبيرة ويميزها عن غيرها من أدوات البحث والاسترجاع.

وتعتمد محركات البحث مثل جوجل وياهو وغيرهما في تصميم واجهات البحث على إتاحة نمط متميز من أيقونات البحث يطلق عليها البحث العمودي (Vertical Search (Iskold, 2006 وتشمل ما يلي:

i. بحث الويب العادي Regular web search وهو أكثر أنماط البحث شهرة وانتشاراً واستخداماً من جانب الباحثين والذي يعتمد على بحث كشافات محركات البحث بصرف النظر عن نوع صفحة الويب سواء كانت متاحة في شكل نص تم إعداده باستخدام لغة تكويد النصوص الفائقة أو غيرها من أشكال الوثائق التي يمكن إتاحتها على الخط المباشر مثل PDFs أو وثائق (Microsoft Office Word, Excel, Power Point, …etc).

ii. بحث الأخبار News Search والذي يمكن من خلاله بحث المواقع الإخبارية فقط للصحف والمجلات ووكالات الأنباء وعادة ما يتم ترتيب النتائج المسترجعة من هذه المواقع تاريخياً بناء على تاريخ الخبر أو الموضوع. فمثلاً إذا كان أحد الباحثين يريد معلومات عن مباراة كرة قدم فسيتم عرض المواقع مرتبة من الأحدث إلى الأقدم.

iii. بحث الصور Image search وتستخدم لبحث الصور التي تم اكتشافها أثناء عمليات حصاد مواقع الويب من خلال الزواحف، وعادة ما يتم تكشيف الصور باستخدام أسماء ملفات الصور image's filename والنصوص المحيطة بالصورة، كما تسعى محركات البحث إلى تطبيق تكنولوجيا الذكاء الاصطناعي كمحاولة لفهم واكتشاف مضمون الصورة ولكن هذه العملية مازالت تسير ببطء. فعلى سبيل المثال يستطيع محرك البحث جوجل الآن فصل صور الوجوه ورسم خطوط من صور أخرى.

iv. بحث الفيديو Video Search ويتم الاعتماد فيه على بحث النصوص المصاحبة لملف الفيديو. ويعتمد دقة البحث في ملفات الفيديو والصور على قيام معدي الصور وملفات الفيديو بوصفها وصفاً دقيقاً سواء من خلال أسماء الملفات أو الميتاداتا أو النصوص المحيطة بهذه الملفات.

توجد أنماط أخرى من أنماط البحث تتضمنها واجهات التعامل تشمل إمكانية بحث المدونات والمجموعات الإخبارية وبحث الإنتاج الفكري العلمي مثل Scholar Search. كما تقوم محركات البحث أحياناً بدمج أنواع البحث المختلفة معاً في صفحة نتائج محركات البحث (Mayer, 2007).

10.3.3 البحث الشخصي
Personal Search

يقوم العديد من محركات البحث بإجراء دراسات وتجارب عن الطرق والأساليب التي يمكن أن تراعي سلوك المستفيدين عند التعامل مع أدوات ومحركات البحث بغرض التعرف إلى أفضل مجموعة من نتائج البحث للباحثين على الويب. فعلى سبيل المثال عند البحث عن مصطلح الزواحف فإن الباحث الذي يبحث عن معلومات فنية عن مصطلح الزواحف فإنه يحتاج معلومات عن زواحف محركات البحث وليس فصيلة الزواحف في الكائنات الحية. وتشير الدراسات أيضاً إلى أن ثلث استفسارات المستفيدين هي استفسارات مكررة وفي معظم الأحيان يرجع

المستفيد إلى الصفحة نفسها التي رجع إليها من قبل، لذلك يمكن لمحركات البحث أن تقوم باختيار الصفحات التي استخدمها المستفيد سابقاً وعرضها في قمة صفحة النتائج المسترجعة وذلك عندما يقوم المستفيد بإدخال مصطلحات الاستفسار نفسها (Teevan et. el., 2006).

ويشير الشكل التالي إلى شاشة لصفحة نتائج من محرك البحث جوجل للبحث الشخصي عن البحث في الويكي (الموسوعات الحرة) حيث يمكن للباحث أن يدعم Promote النتائج بحيث يتم دفعها لقمة صفحة النتائج، واستبعاد Remove النتائج الضعيفة من قائمة نتائج البحث، إلى جانب إضافة تعليقات Comments إلى نتائج بعينها. إلا أن جوجل لم يوضح ما إذا كانت المعالجة الشخصية للنتائج سوف تؤثر في نتائج الآخرين أم لا (Dupont & Anderson, 2008).



شكل رقم (10.4) نموذج لصفحة ويب تدعم البحث الشخصي

مع العلم أن خوارزميات الفرز والترتيب في غوغل تتأثر برد فعل المستفيد فيما يعرف بالصلاحية الراجعة Relevance Feedback عند التعامل مع صفحة النتائج، حيث يتم دفع الصفحات التي يكثر الطلب عليها إلى قمة القائمة ويتم دفع الصفحات التي يقل الطلب عليها إلى ذيل القائمة.

ويمكن القول بإيجاز إن الإنجاز الذي حققته محركات البحث كأداة تساعد على بحث ملايين الصفحات والمواقع المتاحة على الويب في أقل من ثانية تطور كبير وغير مسبوق في آليات البحث والاسترجاع. فكما رأينا فإن محركات البحث لا تقوم ببحث الويب نفسها وإنما تقوم ببحث نسخ من صفحات الويب يتم تجميعها من خلال الزواحف التي تقوم بحصاد صفحات الويب. ويتم تكشيف النتائج في قواعد بيانات محركات البحث التي تتولى ترتيب صفحات الويب بناء على مجموعة من المعاملات (العناوين، تردد المصطلحات، حجم الخط وشكل العرض.. الخ)، إضافة إلى مستوى أهميتها في شكل الويب من خلال تحليل علاقتها بالصفحات الأخرى على الويب. وتجدر الإشارة إلى أنه توجد منافسة بين محركات البحث على عرض أكثر مجموعة نتائج صلاحية للبحث، حيث تسعى كل المحركات إلى تطوير أدئها باستمرار للوصول إلى أفضل أساليب الفرز والترتيب. وكما تتنافس محركات البحث على عرض أفضل نتائج وتطوير مستوى الصلاحية، فإن مواقع الويب تتنافس أيضاً في استخدام أفضل أساليب الترقية لكي يتم عرضها كأول نتيجة في قائمة النتائج المسترجعة.

10.3.4 ملامح البحث في المحركات

تتيح معظم محركات البحث أساليب عدة للبحث عن صفحات ومواقع الويب:

- **البحث البسيط Simple Search**

تعد هذه الطريقة أبسط أساليب البحث وأكثرها سرعة، حيث يتم من خلالها إجراء البحث بكلمة واحدة أو جملة كاملة. وتتم كتابة الكلمة أو الجملة المطلوب البحث عنها في صندوق البحث دون وجود أي روابط تحدد العلاقات بين كلمات البحث. وقد أثبتت الدراسات المتعلقة بتحليل استفسارات المستفيدين أن هذا النمط

من أنماط البحث هو أكثر الأساليب التي يميل المستفيدون إلى استخدامها نظراً لسهولة وسرعة صياغة العبارات البحثية، فهو لا يحتاج من المستفيد أي خبرة مسبقة في عمليات البحث والاسترجاع، هذا إضافة إلى أنه أسرع أنماط البحث، حيث لا يحتاج الباحث إلى بناء طريقة بحث تحدد العلاقات بين كلمات الاستفسار أو الانتقال من الشاشة الرئيسة إلى شاشات أخرى لإجراء عملية البحث. ولكي يستطيع الباحث أن يحقق أعلى معدلات الدقة في البحث باستخدام هذا الأسلوب يجب إتباع التعليمات التالية:

- **استخدام مصطلحات محددة Use Specific Terms**

فكلما كانت المصطلحات المستخدمة في عملية البحث دالة ومستخدمة من جانب المتخصصين في المجالات الموضوعية للدلالة على موضوع البحث، كان من السهل الوصول إلى المعلومات المطلوبة، نظراً لأن معظم محركات البحث تعتمد على تكشيف الكلمات المستخدمة في الصفحات. وهي عادة ما تتضمن المصطلحات السائدة بين المتخصصين.

فعلى سبيل المثال إذا كان الباحث يريد معلومات عن جراحات زراعة الأعضاء Origin Transplant Surgery فمن الأفضل أن يكتب المصطلح كاملاً دون استبعاد أي مفهوم من المفاهيم الثلاثة. فالبعض مثلاً قد يبحث عن هذا الموضوع باستخدام Origin Transplant ومن الواضح أن هذه العبارة البحثية غير كاملة، حيث يمكن أن يسترجع مواد لا علاقة لها بالعمليات الجراحية نظراً لأن المصطلح Surgery غير موجود ضمن مصطلحات الاستفسار. وربما يكون من الأفضل أن تبحث عن المشكلة التي تريد حلها على وجه الدقة باستخدام صيغة السؤال مثل: How to install a memory card in PC. ولعل أكثر الأساليب كفاءة في مثل هذه الحالات هي البحث باستخدام صيغة الجملة أو ما يعرف بالـ Phrase Search والذي سنتناوله بمزيد من التفصيل فيما يلي. مع العلم أن أفضل أساليب البحث كما ذكرنا من قبل هو استخدام أحد استراتيجيات البحث التي سبق عرضها وفقاً للحالة وطبيعة الاستفسار الذي يسعى المستفيد إلى معالجته.

- **استخدام علامة الجمع (+)**

في بعض الحالات قد تكون في حاجة إلى التأكد من أن محرك البحث سوف يسترجع صفحات تتضمن كل الكلمات التي اشتملت عليها صيغة البحث أو أن تكون أحد هذه الكلمات لا يمكن الاستغناء عنها في الصفحات المسترجعة. وفي هذه الحالة تتيح معظم محركات البحث إمكانية وضع علامة + قبل الكلمات المهمة، بالتالي لا يسترجع محرك البحث أي صفحة إلا إذا كانت تتضمن هذه الكلمة.

فعلى سبيل المثال قد تحتاج إلى استرجاع صفحة تتضمن معلومات عن The role of Naser in the preparation for 1973 war (دور جمال عبد الناصر في التحضير لحرب أكتوبر) في هذه الحالة لا يمكن استرجاع أي صفحة لا تتضمن جمال عبد الناصر وحرب أكتوبر بالتالي تكون الصيغة الملاءمة للبحث كما يلي:

The Role of +Naser in the preparation for +1973 +War

بالتالي لا بد أن يقوم محرك البحث باسترجاع صفحات تتضمن كلاً من عبد الناصر وحرب 1973. ومن الممكن أن يسترجع صفحات تتضمن بقية كلمات الاستفسار ولكن محرك البحث سوف يعطي أهمية أكبر لكل من الصفحات التي تتضمن كلاً من ناصر وحرب 1973.

مثال آخر : +Windows 2010+bugs

سوف يقوم محرك البحث باسترجاع الصفحات التي تتضمن هذه المصطلحات الثلاثة في الصفحة نفسها مع إعطاء أهمية أكبر للمصطلحات windows، bug وإعطاء أهمية أقل للمصطلح 2010 ويستبعد أي صفحة لا تتضمن أي من هذه المصطلحات.

وعادة ما يكون استخدام علامة الجمع مفيداً عندما تكون النتائج المسترجعة من البحث البسيط كبيرة جداً ولا يمكن للمستفيد الاطلاع عليها جميعاً في هذه الحالة يكون من المفيد تحديد المصطلحات المحورية والتركيز عليها في البحث من خلال وضع علامة الجمع قبلها، مما يساعد على تضييق نطاق البحث واسترجاع عدد أقل من النتائج التي يسترجعها البحث البسيط.

- **استخدام علامة الطرح (-)**

قد يحتاج المستفيد إلى البحث عن موضوع مع استبعاد جانب معين من جوانب هذا الموضوع أو مصطلح معين من المصطلحات المرتبطة بهذا الموضوع. على سبيل المثال، تخيل أنك تحتاج إلى معلومات عن Bill Clinton وعندما أجريت البحث بالمصطلح Bill Clinton وجدت عدداً كبيراً جداً من الصفحات تتناول قضية Monica Lewinsky وأنت لست مهتماً بهذه القضية في هذه الحالة سوف تكون في حاجة إلى استبعاد كل الصفحات التي تتناول Monica Lewinsky من البحث. من ثم تكون علامة الطرح في هذه الحالة ذات أهمية كبيرة، ويكون البحث كما يلي:

Bill Clinton -Monica -Lewinsky+

بالتالي سوف يقوم محرك البحث باسترجاع كل الصفحات التي تتضمن المصطلح بيل كلينتون مع استبعاد أي صفحة من ضمن الصفحات التي تعالج كلينتون قد تعرضت للمصطلح مونيكا.

مثال آخر قد يكون المستفيد في حاجة إلى استرجاع معلومات عن ويندوز 10 ولكن عند إجراء البحث وجد العديد من الصفحات التي تتناول Windows 8 أو Windows 7 بالتالي يكون المستفيد في حاجة إلى استبعاد هذه الصفحات من خلال استخدام الاستراتيجية التالية:

Windows 10 – Windows 7 -Windows 8+

بالتالي يمكن القول إن علامة الطرح مفيدة بصفة عامة في تركيز البحث على الجانب الأكثر أهمية واستبعاد الجوانب الهامشية، خاصة إذا كانت هذه الجوانب تسترجع عدداً كبيراً من الصفحات غير مرتبطة بموضوع البحث الأصلي أو باحتياجات المستفيد الأساسية.

- **استخدام علامة التنصيص « »**

لقد تعلمنا الآن كيف يمكن أن نجمع النتائج ونطرحها من خلال استخدام علامات

الجمع والطرح. والآن سوف نحاول إلقاء الضوء على عملية الضرب في محركات البحث. وتتم عملية الضرب في مجال استرجاع المعلومات من خلال استخدام علامة التنصيص، حيث يتم وضع المصطلحات في شكل جملة بين علامة تنصيص فيما يعرف بالبحث باستخدام الجمل Phrase Searching. ويعد هذا الأسلوب من أفضل أساليب البحث خاصة إذا كانت مصطلحات البحث يمكن صياغتها في شكل جملة. فعلى سبيل المثال في موضوع البحث Origin Transplant Surgery نجد أن الطريقة المثالية لصياغة هذا الاستفسار هي وضع كلماته بين علامة تنصيص، مما يعنى أن النتائج التي سوف تسترجع لا بد أن تشتمل على هذه الجملة كما وردت في استراتيجية البحث.

مثال: «Origin Transplant Surgery»

مثال آخر: «Search Engines Tutorials»

في هذه الحالة سوف يسترجع محرك البحث كل النتائج التي تشتمل على كل هذه الكلمات ويرتب النتائج حسب عدد مرات تكرار الجملة، ولكن ليس معنى ذلك أن نتائج البحث سوف تقتصر على هذه الجملة فقط ولكن قد يسترجع محرك البحث بعض النتائج التي تشتمل على كلمتين متقاربتين والثالثة قد ترد في أي مكان آخر أو ربما يسترجع محرك البحث بعض النتائج التي تشتمل على هذه الكلمات الثلاث ولكنها غير متقاربة، ولكن هذه النتائج عادة ما ترد في ذيل قائمة النتائج المسترجعة.

والخلاصة أن إجراء البحث باستخدام الجملة يساعد على الوصول إلى نتائج تشتمل على كلمات الاستفسار كما تم إدخالها في صندوق البحث، وفي ترتيبها نفسه، وذلك من خلال وضع علامات التنصيص حول كلمات الاستفسار.

وتجدر الإشارة إلى أن الاتجاه العام في محركات البحث هو استخدام المعامل OR في الربط بين المصطلحات عند البحث، بينما يرتب الوثائق المسترجعة باستخدام المعامل AND كخط أول للترتيب يليه المعامل OR كخط ثانٍ في الترتيب.

- **المزج بين العلامات Operators Combining**

من الممكن أن نحتاج في بعض الأحيان إلى المزج بين أكثر من علامة من

علامات البحث مثل المزج بين الجمع والطرح والضرب. فعلى سبيل المثال قد يحتاج المستفيد إلى البحث عن فضائح بيل كلينتون مع استبعاد فضيحة مونيكا، يمكن إجراء البحث كما يلي: +Bill Clinton Scandals -Monica Lewinsky.

في هذه الحالة سوف يسترجع محرك البحث كل فضائح بيل كلينتون مع استبعاد فضيحة مونيكا من نتائج البحث، أو ربما يحتاج إلى كل ما يتعلق بسياسة أمريكا تجاه الشرق الأوسط مع استبعاد كل ما يتعلق بالصراع العربي الإسرائيلي: USA role in Middle East -Israel.

في هذه الحالة سوف يسترجع محرك البحث كل الصفحات التي تتناول دور أمريكا في الشرق الأوسط مع استبعاد كل ما يتعلق بقضية الصراع العربي الإسرائيلي.

مثال آخر: «تنظيم المعلومات» +الفهرسة +مارك 21 - الميتاداتا

في هذا المثال يحتاج الباحث إلى كل ما يتعلق بالمصطلح «تنظيم المعلومات» كجملة على أن يكون موضوع الفهرسة ومارك 21 مصطلحات أساسية في قائمة النتائج المسترجعة مع استبعاد أي وثيقة تتعامل مع الميتاداتا.

مثال آخر: قد يحتاج المستفيد إلى استرجاع صفحات عن عمليات زرع الأعضاء مع التركيز على زراعة الكبد واستبعاد عمليات زرع الكلى.

الاستراتيجية: Origin Transplant Surgery +Lever Transplant -Kidney+

ومن الجدير بالذكر أن معظم أوامر المنطق البوليني -التي تتيح إمكانية البحث بالكلمات الدالة باستخدام معاملات الربط البوليني AND / OR / NOT- أو البحث بالجمل الكاملة أو البحث التجاوري proximity Search أو إمكانيات البتر Truncation والجذع Stemming كانت تستخدم لفترة طويلة في نظم الاسترجاع التقليدية مثل قواعد البيانات الببليوجرافية ولكنها كانت في غاية الصعوبة بالنسبة للمستفيد العادي مما اضطر القائمين على نظم البحث والاسترجاع إلى الاعتماد على الباحثين المتخصصين لإجراء البحوث للمستفيدين، فيما عُرف بوسيط البحث Search Intermediate، إلا أن محركات

البحث استطاعت التغلب على هذه المشكلة من خلال استخدام علامات أكثر سهولة في بناء استراتيجيات البحث والتي تمثلت في علامات الجمع والطرح والضرب.

ويوضح الجدول التالي الإمكانيات التي توفرها مجموعة من محركات البحث العالمية ومدى قدرتها على استخدام أساليب البحث السابق عرضها:

جدول (10.2) معاملات البحث في محركات البحث ودلالاتها

العلامة	دلالتها
+	لا بد من وجود مصطلح البحث في الصفحات المسترجعة
–	استبعاد الصفحات التي تتضمن المصطلحات التي تلي علامة الطرح
« »	استرجاع الصفحات التي تتضمن الجملة بنفس ترتيب وصياغة المصطلحات

1. البحث المعقد باستخدام معاملات المنطق البوليني:

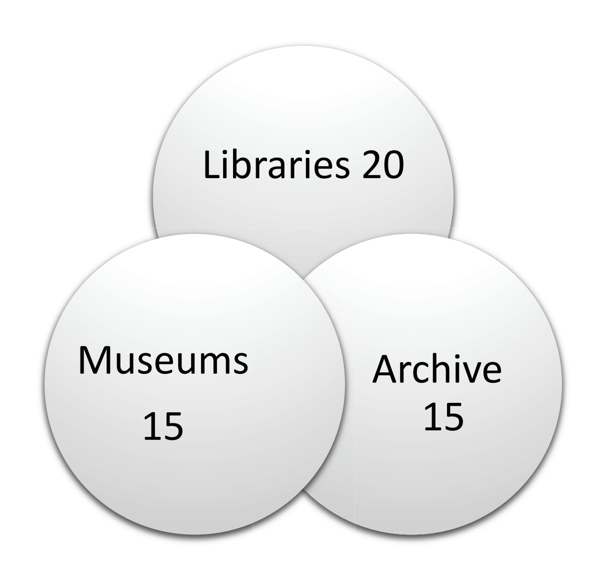
على الرغم من صعوبة البحث بالمنطق البوليني خاصة عندما تكون استفسارات المستفيدين معقدة وطويلة، إلا أن دراسات سلوكيات المستفيدين عند تعاملهم مع محركات البحث أثبتت أن المستفيدين يميلون إلى استخدام عدد قليل من المصطلحات في عمليات البحث والاسترجاع من الشبكة العنكبوتية. فقد أوضحت تحليلات استفسارات المستفيدين على الويب أن متوسط عدد المصطلحات يبلغ 2.4 مصطلح، بينما متوسط عدد المصطلحات في نظم الاسترجاع التقليدية بلغ من 12 إلى 15 مصطلح. وقد أعطى ذلك الفرصة لمحركات البحث لبناء أساليب بحث تعتمد على استخدام المنطق البوليني. وسوف نستعرض فيما يلي العلامات المستخدمة في البحث البوليني على الشبكة العنكبوتية وطريقة الربط بين المصطلحات مع التقيد باستخدام عدد قليل من المصطلحات.

سبقت الإشارة إلى أن عمليات البحث البوليني توظف ثلاثة روابط أساسية للربط بين المصطلحات هي AND, OR, NOT ولا تختلف هذه العلامات في دلالتها كثيراً عن دلالة علامات الجمع والطرح والضرب. ويغطي هذا الجزء طريقة معالجة أوامر

المنطق البوليني من خلال محركات البحث على افتراض أن القارئ قد استوعب الأساليب السابقة والتي سوف تساعده كثيراً على استيعاب ما يلي:

- **المعامل أو - OR**

ويستخدم هذا المعامل للدلالة لتوسيع نطاق البحث عن المفاهيم المتشابهة بمعنى أو ـ أي حيث يعني استرجاع الصفحات التي يظهر فيها أي من المصطلحات الواردة في استراتيجية البحث. بمعنى إذا كان لدينا استراتيجية بحث مكونة من ثلاثة مصطلحات كما يلي: Libraries OR Archives OR Museums



شكل رقم (10.5) استخدام معامل الربط البوليني OR في البحث عن المعلومات

سوف يقوم محرك البحث باسترجاع كل الصفحات التي تتضمن أي مصطلح من المصطلحات الثلاثة، فإذا كانت الصفحات التي تتضمن المصطلحات الثلاثة السابقة موزعة كما يلي:

مع مراعاة أن بعض الصفحات قد تعالج أكثر من موضوع في الوقت نفسه، هذه الصفحات في هذه الحالة تعد مكررات لا بد من استبعادها فمثلاً:

Libraries And Archives 4 Pages

Libraries And Museum 3 Pages

Archives And Museums 4 pages

Libraries And Archives And Museums 2 page

يكـون عـدد الوثائـق المسـترجعة فـي هـذه الحالـة يشـتمل علـى (16 + 11 + 10= 37 صفحـة) معنـى ذلـك أن هنـاك 13 وثيقـة تكـرر بهـا مصطلحـان ووثيقتـان فقـط تكـرر بهمـا المصطلحـات الثلاثـة. بالتالـي يقـوم محـرك البحـث باسـتبعاد كل الوثائـق المكـررة والاحتفـاظ بنسـخة فريـدة مـن أي صفحـة مسـترجعة.

- **المعامل AND**

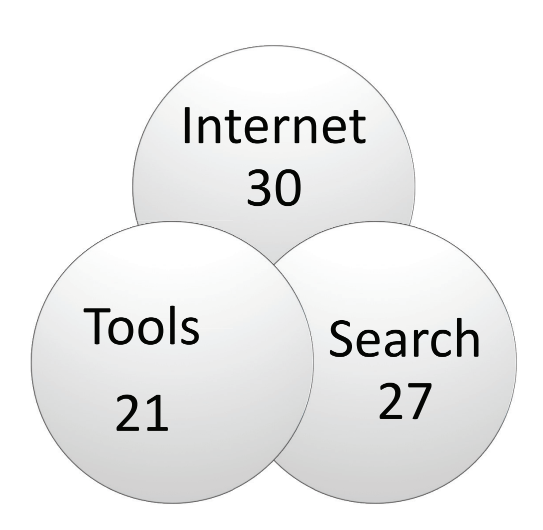
يستخدم هذا المعامل مع المفاهيم المتنوعة في دلالتها لتحقيق الربط بينها، ويعني اسـترجاع كل الصفحـات التـي تتضمـن جميـع المصطلحـات الـواردة فـي اسـتراتيجية البحث معاً، بحيث إذا كان أي من الصفحات لـم يـرد فيها أي من المصطلحات المحددة في استراتيجية البحث يقوم محرك البحث باستبعادها من قائمة النتائج المسترجعة.

مثال: Globalization AND Economic AND Developing Countries

تشـير هـذه الاسـتراتيجية إلـى ضـرورة أن تتضمـن كل الصفحـات المسـترجعة علـى كل المصطلحـات الـواردة فـي اسـتراتيجية البحـث. بمعنـى أن تعالـج كل الصفحـات المسـترجعة موضـوع العولمـة والاقتصـاد فـي الـدول الناميـة.

مثال آخر: Internet AND Search AND Tools

تشـير هـذه الاسـتراتيجية إلـى أن كل الصفحـات المسـترجعة لا بـد أن تتضمـن كل المصطلحات الـواردة في استراتيجية البحث. بالتالي لكي تسترجع أي صفحة لا بـد أن تعالـج موضـوع الإنترنـت والمحـركات والأدوات. وكمـا هـو واضـح مـن الشـكل أنـه نقطـة التقاطـع بيـن المصطلحـات الثلاثـة.



شكل رقم (10.6) استخدام معامل الربط البوليني AND في البحث عن المعلومات

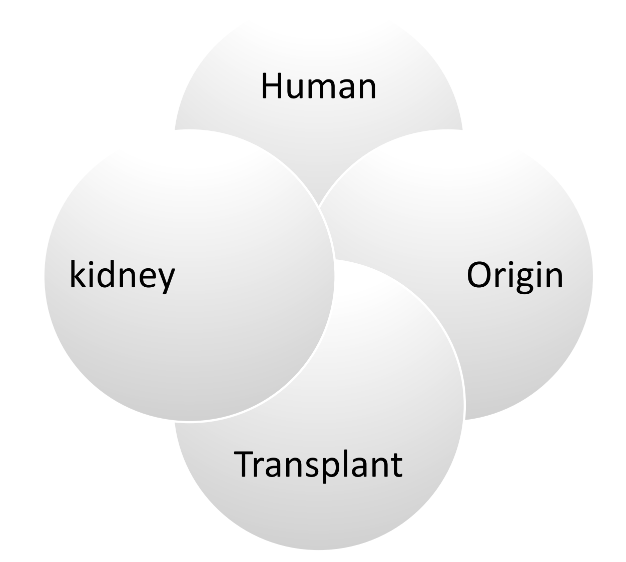
- **المعامل NOT**

يستخدم هذا المعامل مع المفاهيم المرتبطة في الدلالة والتي تشمل علاقات التشابه أو التداخل الهرمي أو التوارث الهرمي، ويعني ماعدا أو باستثناء، ويشير إلى استبعاد الصفحات التي تعالج المصطلحات الواردة بعد المعامل NOT من قائمة النتائج المسترجعة.

مثال: Human AND Origin AND Transplant NOT kidney

تشير هذه الاستراتيجية إلى ضرورة استرجاع كل الصفحات التي تعالج موضوع زراعة الأعضاء للبشر مع ضرورة استبعاد عمليات زراعة الكلى من النتائج المسترجعة.

ونظراً لكفاءة محركات بحث الشبكة العنكبوتية قامت العديد من شركات قواعد البيانات المتاحة على الخط المباشر بشراء محركات لكي تستخدمها كأداة أساسية لبحث قواعد بيانات النصوص الكاملة، ومن أمثلة محركات البحث واسعة الانتشار في هذا المجال محرك البحث Fast، ومحرك البحث Vivisimo حيث يتميز كل منهما بإمكانيات بناء العناقيد (التجميع للمتشابهات وتفريعها) Clustering، والتصنيف إلى فئات Categorization.



شكل رقم (10.7) استخدام معامل الربط البوليني NOT في البحث عن المعلومات

10.3 محركات البحث المتخصصة

حاولت محركات البحث ملاحقة وتتبع التطور والنمو الهائل في الشبكة العنكبوتية ولكن يبدو أن ذلك أمر في غاية الصعوبة، إن لم يكن مستحيلاً، هذا إضافة إلى التنوع الهائل في أنواع الوثائق والحاجة إلى أساليب أكثر فعالية قادرة على التعامل مع الموضوعات ذات الطبيعة الخاصة. وقد دعا ذلك إلى ظهور نوعية جديدة من محركات البحث أطلق عليها محركات البحث المتخصصة Specialized Search Engines للتغلب على مشكلات التغطية التي تواجهها محركات البحث العامة. وتجدر الإشارة إلى أن ظهور لغة التكويد الموسعة Extensible Mark Up Language XML – ساعد على تطور هذا الاتجاه بسرعة كبيرة.

وتعتمد محركات البحث المتخصصة على نوع مميز من الزواحف يطلق عليه الزواحف المركزة Focused Crawler، حيث إنها تركز أثناء عمليات تجميع صفحات الويب على مجموعة من المؤسسات التي لها اهتمامات موضوعية تدخل في

نطاق التخصص الموضوعي لمحرك البحث المتخصص، فتقوم بتتبع خوادم تلك المؤسسات وتجميع المواقع والصفحات التي تشملها تلك الخوادم إضافة إلى متابعة الصفحات والمواقع المرتبطة بها (Ester & Kriegel, 2001).

ويمكن تعريف محركات البحث المتخصصة بأنها «تلك المحركات التي تقتصر في عملية التغطية والبحث إما على مجال موضوعي معين أو نطاق جغرافي محدد Domain name أو نوع معين من الملفات مثل الوسائط المتعددة أو الملفات الموسيقية أو الصور.. الخ». وتتنوع المحركات المتخصصة فمنها محركات البحث التي تغطي نطاقاً جغرافياً معيناً Country and Regional search engines – ويمكن الحصول على قائمة شاملة بمحركات البحث المتخصصة في نطاقات جغرافية محددة من خلال الموقع http://www.philb.com/countryse.htm. وتقتصر مجموعة المحركات التي يضمها هذا الموقع على البحث في دول أو أقاليم جغرافية معينة. بمعنى أنه يهتم بتجميع وبحث الخوادم في نطاقات جغرافية محددة.

كما تشمل محركات البحث المتخصصة موضوعياً محركات تغطي موضوعاً محدداً كالطب مثل Medhunt- http://www.hon.ch/MedHunt أو تقتصر على نوع معين من الملفات كالصور سواء الثابتة أو المتحركة أو الصوت فيما يعرف بمحركات بحث الوسائط المتعددة مثل http://www.musicsearcher.com.

ويشير بريس (Price, 2003) إلى أن محركات المتخصصة يمكن تقسيمها لأربع فئات أساسية هي:

1. محركات بحث متخصصة في شكل أو موضوع معين والتي تمثل جزءاً من محركات البحث العامة. وهذا النمط موجود الآن في معظم محركات البحث العامة التي تتيح إمكانية بحث الملفات ذات الطبيعة الخاصة مثل الصور وملفات الفيديو، من خلال واجهة تعامل خاصة ومنها ما يتيح واجهات تعامل خاصة للأطفال تتمتع بإمكانيات تساعد على تنقية Filtering الصفحات المسترجعة من المواد غير الصالحة لهذه الفئات العمرية. ومن أمثلة هذه النوعية ما يلي:

Google Images (images only)
http://images.google.com

يعمـل هـذا المحـرك كجـزء مـن محـرك البحـث Google وهـو متخصـص فـي بحـث الصـور المتاحـة على شـبكة الإنترنـت.

Lycos Pictures and Sounds -
/http://multimedia.lycos.com

ويعمـل هـذا المحـرك أيضـاً كجـزء مـن محـرك البحـث Lycos وهـو متخصـص فـي بحـث الملفـات الصوتيـة والصـور.

Ask Jeeves For Kids
/http://www.ajkids.com

يقـوم بالبحـث عـن المـواد الخاصـة بالأطفـال مثـل أفـلام الكارتـون والصـور والمـواد التعليميـة وهـو أيضـاً كجـزء مـن المحـرك Ask Jeeves.

Yahooligans
/http://www.yahooligans.com

تـم تصميـم هـذا المحـرك كجزء مـن محرك البحث Yahoo وهـو متخصـص في مـواد الأطفـال مـن سـن 7 إلـى 12 عامـاً، وهـو مـن أقـدم محـركات البحـث المتخصصـة للأطفـال وقـد تـم إنشـاؤه فـي مـارس 1996.

2. محـركات بحـث متخصصـة قائمـة بذاتهـا ولهـا برامـج خاصـة للزحـف والتكشـيف والبحـث. وتركـز فـي تغطيتهـا علـى مجـالات موضوعيـة معينـة أو أنـواع معينـة مـن الملفـات مثـل محـركات بحـث الوسـائط المتعـددة. ومـن أبـرز أمثلـة هـذه النوعيـة مـن المحـركات مـا يلـي:

Health On The Net: MedHunt

/http://www.hon.ch/MedHunt

MedicineNet.com

http://www.medicinenet.com/script/main/hp.asp

وهي محركات بحث متخصصة في مصادر المعلومات الطبية التي يشارك بها أكثر من 500 طبيب ومتخصص من دول مختلفة على رأسها أمريكا وكندا.

3. محركات بحث تستخدم في البحث داخل الأدلة الموضوعية العامة حيث يمكن من خلالها إدخال مصطلحات في صندوق بحث يشبه صندوق البحث في المحركات التقليدية ثم تستخدم تلك المحركات في بحث الدليل الموضوعي. وهذا هو النمط السائد في معظم أدلة البحث العربية التي تقدم إمكانيات للبحث مثل فارس نت والرادادي والبوابة العربية وغيرها.

4. محركات بحث متخصصة صُممت خصيصاً لكي تستخدم في بحث مواقع محددة تشتمل على قواعد بيانات خلفية يطلق عليها صفحات الخوادم النشطة Active Server Page. وتتولى هذه المحركات تلقي استفسارات المستفيدين وتحويلها إلى قواعد البيانات حتى يمكن الحصول على الإجابات وهو نمط سائد في كثير من مواقع الشركات والمؤسسات التي لها بيانات خاصة. وتجدر الإشارة إلى أن هذه الصفحات عادة ما يطلق عليها الصفحات الديناميكية Dynamic Pages. مثال محرك بحث شركة Amazon لتجارة الكتب http://www.amazon.com هو متحرك بحث متخصص للبحث في قاعدة بيانات شركة Amazon للتجارة في مصادر المعلومات من كتب وغيرها.

10.4 ما وراء المحركات
Meta Search Engines

تعد ما وراء المحركات واحدة من أحدث أدوات بحث واسترجاع مصادر المعلومات المتاحة على الشبكة العنكبوتية في الوقت الحالي. وتقوم هذه المحركات بصفة عامة بتلقي استفسارات المستفيدين وإرسالها إلى مجموعة منتقاة من محركات البحث المستقلة. ثم تتلقى النتائج من هذه المحركات وتقوم بدمجها ومعالجتها ثم فرزها في قائمة مرتبة وفقاً لخوارزميات الدمج والترتيب ـ-Merging Algorithms

هـذا إضافـة إلـى بعـض العمليـات الأخـرى مثـل تحليـل الاستفسـارات وترجمتهـا لكـي تتوافـق مـع إمكانيـات البحـث المختلفـة للمحـركات المشـاركة فـي النظـام، ولكـي تسـتفيد أيضـاً مـن القيمـة المضافـة لعمليـات التشـغيل التبادلـي ـIntroperability- التـي توفرهـا خوارزميـات الدمـج والترتيـب (Yang, X. & Zhang, 2000).

وتتمثل المشكلة الرئيسة في بناء ما وراء محركات في ثلاثة تحديات أساسية هي:

- اختيـار محـركات البحـث المسـتقلة وتجميعهـا فـي قائمـة موحـدة وترتيبهـا وفقـاً لأولويـات الدمـج.
- دمج النتائج المسترجعة.
- ترتيب وفرز النتائج المسترجعة.

وفـي مـا يلـي عـرض للأسـس والمعاييـر المسـتخدمة فـي بنـاء مـا وراء المحـركات فـي كل مرحلـة مـن المراحـل الثـلاث السـابقة:

10.4.1 اختيار محركات البحث المستقلة وتجميعها في قائمة موحدة وترتيبها وفقاً لأولويات الدمج

تعـرف هـذه العمليـة فـي الإنتـاج الفكـري المتخصـص فـي مجـال اسـترجاع المعلومـات بعمليـة اختيـار وفـرز قواعـد البيانـات Database Selection and Ranking، حيـث يقـوم الفريـق فـي هـذه المرحلـة بتجميـع قوائـم شـاملة بمحـركات البحـث المسـتقلة للاختيـار مـن بينهـا وفقـاً لأحـد المعاييـر التاليـة (Mohamed, 2004).

أ. حجم التغطية في محركات البحث المستقلة
Individual Search Engines Coverage

فـي هـذه الحالـة يقـوم فريـق العمـل بتجميـع قائمـة شـاملة بأشـهر محـركات البحـث المتاحـة وأكثرهـا شـمولاً مـن حيـث عـدد الصفحـات التـي تـم تكشـيفها والمتاحـة فعليـاً للبحـث والاسـترجاع. ثـم يقـوم بالمقارنـة بيـن قواعـد البيانـات وذلـك عـن طريـق تشـغيل

برنامج للفرز Merging Program حيث يقوم هذا البرنامج بفرز قواعد البيانات وترتيبها تنازلياً من الأكثر شمولاً إلى الأقل فالأقل. ونظراً لأن محركات البحث المستقلة تُنوع في تغطيتها لمصادر المعلومات المتاحة على شبكة الإنترنت من حيث نوع صفحات المعلومات (مثل صفحات الويب، صفحات البي دي إف، صفحات الأوفيس، أو قواعد البيانات، الصور، الفيديوهات.. الخ) فتتم المقارنة بين هذه الأنواع المختلفة لترتيب المحركات وفقاً للاحتياجات الأساسية لما وراء المحركات وليس السياسات المتبعة في المحركات المستقلة. وتجدر الإشارة هنا إلى أنه توجد مصادر متعددة على الشبكة العالمية توفر إحصاءات دقيقة عن معدلات التغطية في محركات البحث المستقلة. ومن أبرز هذه المصادر:

Search Engine Watch
http://searchenginewatch.com
Search Engine List
http://www.thesearchenginelist.com
Search Engine Market Share Worldwide | StatCounter Global Stats
http://gs.statcounter.com/search-engine-market-share

II. معدلات الاستخدام أو الاستفسار Query Load

في هذه الحالة يتم تحديد عدد الاستفسارات التي توجه إلى كل محرك بحث على حدة وترتيبها من المحرك الأكثر استفساراً إلى الأقل استفساراً. كما أن بعض ما وراء المحركات تأخذ في الاعتبار نسبة الاستفسارات الناجحة إلى نسبة الاستفسارات الفاشلة. ويمكن الحصول على هذه الإحصاءات من خلال تحليل ملف الاستفسارات أو ما يعرف بملف اللوج Log File في كل محرك مستقل على حدة. لكن من عيوب هذه الطريقة أنها تتطلب قدراً كبيراً من التعاون من المحركات المستقلة، وهو أمر غير مرغوب فيه في تلك البيئة، نظراً للطبيعة التنافسية الشديدة التي تحكم هذا المجال. فالحصول على هذه الملفات قد يؤدي إلى الكشف عن أساليب تحليل الاستفسارات والخوارزميات المستخدمة في عمليات التكشيف والاسترجاع. هذا وإن كانت هذه الأمور من السهل الكشف عنها من خلال الفحص والتحليل الدقيق للنتائج المسترجعة والأساليب المفضلة لدى هذه المحركات في بناء استراتيجيات

البحث. ولعل أبرز نماذج التعاون في هذا المجال هو ما قدمته محركات البحث المستقلة (Excite, AltaVista and Ask Jeeves) – للباحثين من ملفات بغرض التحليل والدراسة، للتعرف إلى طبيعة الاستفسارات الموجهة إلى هذه المحركات. ومن أمثلة الدراسات التي تناولت محركات البحث المستقلة بالفحص والتحليل ما يلي (Mohamed, 2004; Meng & Lui, 2002).

III. وقت الاستجابة Response Time

يتم قياس متوسط الوقت الذي يستغرقه كل محرك على حدة في إجراء البحث واستعراض النتائج، ثم يتم ترتيب المحركات وفقاً لسرعة الاستجابة من الأكفأ إلى الأقل كفاءة. هذا وإن كان الفارق بين محركات البحث من حيث وقت الاستجابة هو فارق غير محسوس، إلا أن مؤشر وقت الاستجابة عامل في غاية الأهمية بالنسبة لمطوري ما وراء المحركات، نظراً لما تتطلبه العملية من إجراء البحث في أكثر من محرك مستقل. بالتالي فإن سرعة المحركات المستقلة تؤثر بالتبعية على سرعة ما وراء المحركات. وهذه الطريقة سوف تضمن كفاءة عالية من حيث سرعة الاستجابة ولكنها لا يمكن أن تضمن بأي حال من الأحوال كفاءة وفعالية المواد المسترجعة.

IV. تقييم النتائج المسترجعة من المحركات المستقلة Individual Search Engines Results Evaluation

ويشمل التقييم ثلاثة معايير أساسية من مقاييس التقييم في مجال استرجاع المعلومات وهي:

الاستدعاء والدقة والترتيب أو الفرز. ويوجد العديد من الدراسات التي قارنت بين محركات البحث من حيث دقة النتائج المسترجعة. وتتسم هذه الدراسات بالمقارنة بين محركات البحث في بيئتها الطبيعية من حيث مقومات البحث والمجموعات وطبيعة الاستفسارات. ويعرف هذا الاتجاه في الأدبيات بالاتجاه العملي Operational Approach. كما يوجد نوع آخر من الدراسات تولى المقارنة بين محركات البحث المستقلة عن طريق فصل عناصر المقارنة لتجربتها في المعمل. ويعرف هذا الاتجاه بالاتجاه المعملي Laboratory Approach.

حيث تتم التجارب على عناصر معينة في محركات البحث دون العناصر الأخرى للتعرف على مدى تأثيرها في كفاءة ودقة الاسترجاع (Yanh & Zang, 2000).

10.4.2 دمج النتائج المسترجعة
Fusing or Combining Search Results

توجد أربع طرق أساسية لدمج البيانات معروفة ومستخدمة في مجال استرجاع المعلومات. وهذه الطرق هي:

I. دمج النتائج المسترجعة وفقاً لاستراتيجيات بحث متنوعة
Fusing Different Search Strategies

وتعتمد هذه الطريقة على التنويع في طريقة بناء استراتيجية البحث لنفس موضوع الاستفسار، حيث يتم توجيه هذه الاستراتيجيات المتنوعة للمحرك نفسه. ثم يتم دمج النتائج المسترجعة بعد استبعاد النتائج المكررة Overlapped Results. بمعنى أنه عند توجيه استراتيجيات بحث متنوعة للمحرك نفسه يمكن الحصول على نتائج متنوعة ولكنها تدور في مجملها حول موضوع البحث الأساسي مع وجود قدر كبير من التداخل والتكرار بين نتائج هذه الاستراتيجيات المتنوعة. وقد أثبت كل من سيراسيفيك وكانتور (Saracevic & Kantor, 1998) أن هذه العملية تساعد على استرجاع نتائج مختلفة ولكنها متقاربة، كما أن بعض هذه النتائج تكون صالحة والبعض الآخر يكون غير صالح.

II. دمج النتائج المسترجعة وفقاً لأساليب متنوعة لوزن المصطلحات
Fusing According to Term Weighting Schemes

في هذه الحالة يتم استخدام مجموعة موحدة من الوثائق في بناء قواعد بيانات عدة وفقاً لطرق متنوعة لوزن المصطلحات. ثم يتم توجيه الاستفسار نفسه لكل قاعدة بيانات على حدة، ثم يتم دمج النتائج المسترجعة من قواعد البيانات بعد استبعاد المكررات. وقد أكد لي أن استخدام أكثر من طريقة لوزن المصطلحات يؤدي إلى تحسين كفاءة الاسترجاع (Lee, 1995).

III. دمج النتائج وفقاً لأجزاء الوثائق المكشفة
Data Fusion According to Document Representation

تعتمد هذه الطريقة على التنويع في أجزاء الوثائق المكشفة، حيث يتم إعداد قواعد بيانات مستقلة حسب الجزء المكشف من الوثيقة. فعلى سبيل المثال يتم تكشيف عناوين الوثائق فقط في قاعدة بيانات ويتم تكشيف المستخلصات في قاعدة بيانات أخرى. ويتم إجراء البحث في كل قاعدة بيانات على حدة، ثم تُدمج النتائج المسترجعة بعد استبعاد المكررات، لتحديد مدى تأثير هذه الأجزاء في فعالية الاسترجاع. وقد اكتشف كاتزر وزملاؤه أن إجراء البحث على أجزاء متنوعة من الوثيقة يؤدي إلى استرجاع نتائج بنفس الكفاءة والفعالية، مما يؤدي إلى زيادة معدلات الدقة والاستدعاء عند دمج هذه النتائج (Katzer, et. el., 2982).

IV. دمج النتائج المسترجعة من نظم استرجاع متعددة
Data Fusion According to Multiple Retrieval Systems

في النماذج الثلاثة السابقة يمكن استخدام نظام استرجاع موحد مع التنويع في طرق التكشيف أو بناء استراتيجيات البحث أو أجزاء الوثائق المكشفة. أما في هذا النموذج فيتم التنويع في المصدر بأكمله. حيث يتم الدمج من مصادر متعددة Multiple Sources. وهذا هو النموذج السائد في كل ما وراء المحركات والنظم التي تعتمد على استخدام بروتوكول استرجاع المعلومات Z39.50. ومن الفروق الأساسية أيضاً أن الطرق الثلاث السابقة تُكشف مجموعة موحدة من الوثائق، بينما يعتمد هذا النموذج على مجموعة مختلفة من الوثائق مع وجود قدر من التداخل والتكرار بين هذه المصادر المتنوعة (Mohamed, 2004).

وتجدر الإشارة هنا إلى أنه توجد أربع حالات لمجموعة الوثائق المكشفة تصلح لعملية دمج البيانات. وهذه الحالات هي (Yang & Zhang, 2000):

- **حالة التساوي Equivalent Case**

- وهي الحالة التي تكون فيها الوثائق المكشفة في كل قواعد البيانات واحدة دون أي اختلاف فيما بينها.

- **حالة الاشتمال Inclusion Case**
- وهـي الحالـة التـي تكـون فيهـا إحـدى قواعـد البيانـات شـاملة وقواعـد البيانـات الأخـرى تتضمـن جـزءاً مـن الوثائـق المكشـفة فـي قاعـدة البيانـات الشـاملة.
- **حالة الاختلاف Disjoint Case**
- وهـي الحالـة التـي لا يوجـد فيهـا أي تشـابه بيـن قواعـد البيانـات مـن حيـث مجموعـة الوثائـق المكشـفة.
- **حالة التداخل والتكرار Overlapping Case**
- هـي الحالـة التـي تتداخـل فيهـا قواعـد البيانـات مـن حيـث مجموعـة الوثائـق المكشـفة. وهـذه هـي الحالـة السـائدة في كل مـا وراء المحـركات المتاحـة على شـبكة الإنترنـت.

◄ 10.4.3 فرز وترتيب النتائج المسترجعة Results Merging / Ranking

تعـد هـذه الخطـوة أكثـر الخطـوات أهميـة فـي عمليـة دمـج النتائـج المسـترجعة فـي مـا وراء المحـركات، حيـث إن معظـم هـذه المحـركات عـادة مـا تسـتخدم الوسـائل والأسـاليب نفسـها فـي الخطوتيـن السـابقتين، بينمـا يعـد الأسـلوب المسـتخدم فـي مرحلـة الفـرز والترتيـب هـو العنصـر المميـز لمحـرك عـن الآخـر. وعمومـاً، يوجـد أسـلوبان أساسـيان يسـتخدمان لتحديـد الترتيـب الأمثـل للنتائـج المسـترجعة وهمـا:

- **التحميل والتحليل Downloading and Analyzing**

- **الترتيب وفقاً للافتراضات المنطقية Merging According to Logical Assumptions**

وفـي مـا يلـي عـرض لـكل أسـلوب مـع التركيـز علـى الخوارزميـات المسـتخدمة والأسـاس الـذي بنيـت عليـه:

أ. أسلوب التحميل والتحليل

يعرف هذا الأسلوب في أدبيات استرجاع المعلومات بأسلوب فحص أو تفتيش الوثائقDocuments Fetching. ويعتمد هذا الأسلوب على تحميل الوثائق المسترجعة بأكملها أو أجزاء منها من خادم محرك البحث المستقل إلى خادم ما وراء المحركات. ثم يتم تحليل هذه الوثائق باستخدام وسائل متعددة لعل أشهرها حساب درجة التشابه Similarity Score باستخدام طرق متنوعة لوزن المصطلحات (Term Weighting Schemes (Meng & Liu 2002. وتستخدم درجة التشابه في ترتيب الوثائق حسب ارتباطها بموضوع الاستفسار، وحساب درجة التشابه بين مصطلحات الاستفسار والكلمات المكشفة من الوثيقة. ويوجد العديد من نظم التحميل والتحليل المتاحة حالياً، ولعل أبرزها gGoiss, CORI, and CVV . وتجدر الإشارة هنا إلى أن هذه النظم عادة ما تتضمن خوارزميات للاختيار والتحميل والتحليل والدمج في الوقت نفسه، حيث إنها عادة ما تتضمن كل الوظائف اللازمة لما وراء المحركات.

ولعل أبرز مميزات أسلوب التحميل والتحليل هو الاعتماد على أسلوب موحد في التحليل والترتيب بصرف النظر عن الخوارزميات التي تستخدمها المحركات المستقلة في الترتيب. ولهذا النموذج عيوب عدة، لعل أبرزها:

1. أنه يحتاج إلى وقت طويل لتحميل وتحليل الوثائق وهو ما لا يتناسب مع طبيعة مستخدمي الويب.

2. أنه يتطلب مساحات تخزين كبيرة، حيث يتم تحميل الوثائق المسترجعة على خادم ما وراء المحركات، هذا إضافة إلى خوارزميات التكشيف والبحث والتحليل والفرز.

3. يحتاج هذا النموذج إلى أنظمة استرجاع ذات كفاءة عالية لكي تقوم بعمليات التحليل والترتيب بفاعلية وسرعة، حيث إن عمليات البحث في المحركات المستقلة والتحميل والتحليل وبناء ملفات الوثائق واستبعاد المكررات وبناء القوائم الموحدة، ثم في النهاية استخدام أسلوب موحد لعرض النتائج المسترجعة، كل

هذه العمليات لا بد أن تتم على الهواء(*) On the Fly وهي عمليات معقدة ودقيقة إلى درجة بعيدة. ويصلح هذا النموذج ويعمل بكفاءة عالية في نظم التجميع على الخط المباشر Aggregator Online Systems. وهي النظم التي يقوم فيها المورد بتجميع أكبر عدد ممكن من قواعد البيانات، ويتيحها للاسترجاع على الخط المباشر. بالتالي فإن هذه البيئة تسمح بقدر كبير من التعاون بين قواعد البيانات المستقلة ونظام التجميع. ولعل أبرز نموذج لذلك ما يحدث في أدوات الاكتشاف مثل Summon, EDS, MUSE Discovery وغيرها وهو ما لا يتوافر في بيئة الويب التي تقوم على التنافس الشديد بين محركات البحث.

II. أسلوب الترتيب وفقاً للافتراضات المنطقية Merging Upon Logical Assumptions

يعتمد هذا الأسلوب على استخدام الترتيب الأصلي للوثائق المسترجعة من المحركات المستقلة في إنتاج قائمة موحدة من خلال بناء خوارزميات فرز وترتيب تعتمد على الافتراضات المنطقية وتصلح أن تستخدم في ترتيب الصفحات المسترجعة بالاعتماد على البيانات المتوافرة من المحركات المستقلة عن تريب الصفحات وحجم قاعدة البيانات وأهمية تلك الصفحات بصفة عامة. ومن أبرز الخوارزميات المستخدمة في هذا النموذج:

III. الحشو والإدراج Interleave

تعتمد هذه الطريقة على ترتيب قواعد البيانات ترتيباً تنازلياً وفقاً لمقاييس متعددة، مثل شمول التغطية، دقة الاسترجاع، أو وقت الاستجابة. ثم يتم ترتيب الوثائق وفقاً لترتيب قواعد البيانات، حيث تأتي الوثيقة رقم 1 من قاعدة البيانات رقم 1 في الترتيب رقم 1 في القائمة الموحدة، تليها الوثيقة رقم 1 من قاعدة البيانات رقم 2، ثم الوثيقة رقم 1 من قاعدة البيانات رقم 3، ثم الوثيقة رقم 2 من قاعدة البيانات رقم 1، وهكذا إلى أن يتم الحصول على العدد المرغوب من الوثائق في القائمة الموحدة (Meng & Liu 2002).

ويستند نموذج الحشو والإدراج على افتراض أن الوثيقة المسترجعة من محرك بحث أكثر أهمية ربما تكون أفضل من وثيقة أخرى لها الترتيب نفسه، واسترجعت من محرك آخر أقل أهمية. ومصطلح أهمية هنا يشير إلى موقع محرك البحث في قائمة المحركات المستقلة.

١٧. تحويل أرقام الوثائق إلى رقم تشابه عام
Convert Document Rank to Global Similarity Scores

قام لي بتصميم نموذج لترتيب القوائم النهائية يعرف باتجاه درجة التشابه. ويستخدم هذا النموذج الترتيب الأصلي للصفحات الذي تنتجه المحركات المستقلة من أجل ترتيب قوائمها في إنتاج القائمة الموحدة. ويعتمد هذا النموذج على المعادلة التالية (Lee, 1997).

والافتراض الأساسي هنا أن الوثيقة المسترجعة ضمن عدد أكبر من الوثائق أفضل من وثيقة أخرى لها الترتيب نفسه ومسترجعة ضمن عدد أقل من الوثائق. فعلى سبيل المثال، فإن الوثيقة رقم 1 المسترجعة ضمن ألف وثيقة تعد أفضل من وثيقة رقم 1 ومسترجعة ضمن خمسمئة وثيقة.

ترتيب الوثيقة - 1

درجة التشابه = 1-

عدد الوثائق المسترجعة من المحرك المستقلة

كما قام كل من يونو ولي بإعداد معادلة لتحويل رقم الوثيقة المحلي Local Rank Score إلى رقم تشابه عام Global Similarity Score من خلال تطبيق المعادلة التالية (Yuwono & Lee, 1996).

(*) على الهواء On The Fly تعني أن المستفيد على اتصال مباشر بالخادم الذي يقوم بأداء كل هذه العمليات المذكورة.

نفترض أن لكل استفسار في ترتيب محرك البحث Di هو ri وأن rmin هو ترتيب آخر قاعدة بيانات في القائمة، r هو الترتيب المحلي للوثيقة المسترجعة، g هي درجة التشابه العام. والمعادلة المستخدمة في ترتيب القائمة النهائية:

g = 1- (r – 1) * Fi

حيث إن F هي:

(Fi = (rmin) / (m * ri

وإن m تمثل العدد المرغوب من الوثائق في القائمة النهائية.

فعلى سبيل المثال نفترض وجود قاعدتي بيانات D1 وD2 ونفترض أن ترتيبهم r1 = 0.2 وr2 = 0.5 ونفترض أن العدد الكلي المطلوب من الوثائق هو أربع وثائق، بالتالي فإن:

rmin = 0.2 , F1 = 0.25, F2 = 1, m = 4

ووفقاً للمعادلة فإن الوثائق الثلاث الأولى في D1 سوف يحصلون على درجات تشابه 1, 0.75، 0.5 على التوالي. والوثائق الثلاث الأولى من D2 سوف يحصلون على درجات تشابه 1، 0.9، 0.8 على التوالي. من ثم فإن القائمة النهائية سوف تتضمن ثلاث وثائق من D2 ووثيقة واحدة من D1 هم على التوالي: 1، 1، 0.9، 0.8.

10.4.4 نماذج لما وراء المحركات المتاحة على شبكة الإنترنت

لقد ظهر العديد من أدوات البحث التي تستخدم تقنية ما وراء المحركات خلال الأعوام القليلة الماضية. ويمكن الوصول إلى قوائم شاملة بتجارب بناء ما وراء المحركات من موقع رصد ومشاهدة تطورات محركات البحث:

Search Engine Watch http://searchenginewatch.com

وسوف نستعرض فيما يلي نماذج لأفضل التجارب لبناء ما وراء المحركات.

اشتملت صفحة المعلومات[1] Search Engine Watch في فبراير 2018 على 71 أداة بحث تستخدم تقنية ما وراء المحركات. بعض هذه الأدوات تعرض قائمة

شاملة بمحركات البحث المستقلة المرشحة للبحث مثلStartpage, DuckDuck Go, Dogpile والبعض الآخر لا يعرض المحركات المستقلة المشاركة في ما وراء المحركات مثل, Profusion Excite حيث تستخدم هذه المحركات قالباً عاماً للبحث. ومع ذلك يمكن الوصول إلى القائمة المستخدمة في البحث من خلال خيارات البحث المتقدم Advanced or Customized Search Options.

وبمراجعة أبرز النماذج المتاحة لما وراء المحركات أتضح أن المحرك[2] (Dogpile http://www.dogpile.com) لا يقوم بدمج النتائج المسترجعة، إنما يستعرض نتائج كل محرك مستقل على حدة، بينما يقوم كل من Startpage and Mamma بدمج النتائج من خلال استخدام المكررات في ترتيب القائمة النهائية، حيث يتم الدفع بالوثائق التي تظهر في أكثر من محرك بحث مستقل إلى قمة القائمة. بالتالي فإن الوثيقة التي تظهر في ثلاثة محركات تسبق وثيقة أخرى ظهرت في محركين فقط. وتقوم أداة البحث (MetaCrawler http://www.metacrawler.com/) بجمع درجة تشابه الوثائق المكررة بالتالي تحصل الوثائق المكررة على درجة أعلى من الوثائق الفريدة Unique Documents.

وتعتمد أداة البحث (Profusion (http://www.profusion.com/index.htm على وزن المصطلحات، حيث يتم استخدام كل من درجة التشابه المسترجعة من المحركات المستقلة والدرجة التي حصل عليها محرك البحث المستقل في مرحلة ترتيب المحركات المستقلة. ولكن المشكلة الأساسية في هذه الطريقة أنه ليست كل المحركات المستقلة تسترجع الوثائق مصحوبة بدرجة التشابه، ولكنها تسترجع الوثائق مرتبة فقط دون أي معلومات إضافية عن الدرجة التي حصلت عليها كل وثيقة. بالتالي يتطلب استخدام هذه المعادلة تعاون المحركات المستقلة مع ما وراء المحركات (Callan..Connel, 2001)

أما أداة البحث ميتاجير (MetaGer) http://meta.rrzn.uni-hannover.de فتعتمد على نظام التحليل والتحميل لترتيب القائمة النهائية. حيث تستخدم الترتيب الأصلي للوثائق المسترجعة من المحركات المستقلة إلى جانب تردد المصلحات في عناوين تلك الوثائق، أو ما وراء البيانات Metadata أو ملخص الوثيقة. كما تعتمد أداة البحث Inquiries على نظام التحليل والتحميل، حيث يتم تحميل الوثائق بالكامل على

خادم ما وراء المحركات ثم تحليلها وبناء الكشافات. وتجدر الإشارة هنا إلى أن أداة البحث[1] Inquiries تعتمد على تردد المصلحات إضافة إلى تقارب المصلحات Term Proximity من أجل ترتيب القوائم النهائية.

وتستخدم أدوات ما وراء المحركات بكثافة في مواقع حجز الفنادق وشركات الطيران، حيث تمكن تلك الأدوات من البحث بكفاءة في محركات البحث لشركات الطيران والفنادق لتقديم أفضل عروض الشراء الخاصة بتذاكر الطيران وعروض الفنادق.

10.5 بوابات الويب
Web Portals

يوجد عدد كبير من المصطلحات المستخدمة للدلالة على مفهوم بوابات الويب منها فهارس الإنترنت Internet Catalogs، والمداخل Gateways، والبوابات Portals، والبوابات الموضوعية Subject Portals.. .الخ. وتشير هذه المصطلحات إلى مجموعة الأدوات التي تسعى إلى تنظيم مصادر المعلومات المتاحة من خلال تقسيمات موضوعية شاملة بحيث تشتمل البوابة على جميع أنواع المصادر والخدمات التي يحتاج إليها المستفيدون من خدمات الشبكة العنكبوتية مثل خدمات بريد إلكتروني، دردشة، قوائم خدمات وقوائم بريدية، المواد الإخبارية، أسعار العملات، أحوال الطقس، إلى جانب قوائم موضوعية بمصادر المعلومات المتاحة من خلال البوابة إلى جانب محرك يتيح إمكانية البحث في البوابة. وإلى جانب التنوع في الخدمات التي تقدمها البوابات للمستفيدين منها، نجد أن هذه المواقع عادة ما تتضمن برامج تساعد على تحليل استخدامات المستفيدين Web Usage Analyzer وتساعد على تحليل التوجهات بغرض بناء ملفات سمات المستفيدين User Profiles ويمكن من خلال هذه الملفات التعرف إلى احتياجات المستفيدين والتنبؤ بها بالتالي اختيار

(1) Big Search Engines Index--- http://www.search-engine-index.co.uk

(2) ملحوظة المحرك Dogpile قام بتغيير استراتيجيته للدمج والفرز في شهر يوليو 2005 حيث أصبحت تعتمد على عدد مرات النقر على الصفحة وفتحها في كل محرك مستقل.

المصادر المناسبة لكل مستفيد من المستفيدين من الموقع. ويمكن أن تقوم تلك المواقع باستخدام تكنولوجيا الدفع Pushing Technology إلى المستفيدين من الموقع. ويمكن أن تتم عملية الدفع عبر خدمات البريد الإلكتروني التي توفرها تلك المواقع أو إلى الصفحات الأمامية للمستفيدين من هذه المواقع كما يمكن أن يتم الدفع إلى دوسيهات خاصة للمستفيدين من هذه المواقع.

من ثم فالبوابات عادة ما تيسر لمستخدمي تلك المواقع كل أنواع الخدمات التي يحتاجون إليها بصورة تفاعلية مما يوفر كل احتياجات المستفيد من خدمات ومصادر الشبكة العنكبوتية. وفي مقابل ذلك تسعى البوابات إلى جذب الشركات التي تسعى إلى الإعلان عن منتجاتها وخدماتها لتحقيق الأرباح من خلال تلك المواقع حيث إنه من المعروف أنه كلما زاد عدد مستخدمي الموقع تهافتت الشركات على الإعلان عن خدماتها ومنتجاتها من خلال هذه المواقع (Miller, 2005).

وتنقسم بوابات الويب وفقاً للجمهور الذي تخدمه إلى نوعين أساسيين هما (Yakal, 2005).

◄ 10.5.1 البوابات العامة
General Portals

يقدم هذا النوع من البوابات خدماته لقطاع عريض من المستفيدين من الشبكة العنكبوتية بصرف النظر عن النشاط أو التخصص الموضوعي أو المجال الجغرافي للصفحات التي تغطيها البوابة. وعادة ما توصف هذه النوعية من البوابات بأنها بوابات أفقية Horizontal Portals حيث إنها تعمل على نطاق أفقي سواء من حيث التغطية الموضوعية أي تغطي كل مجالات المعرفة البشرية أو على النطاق الجغرافي أو

(1) Inquiries لم يعد متاحاً على الويب وهو أداة بحث أعدها كل من لورانس وجيل لتحليل معدلات الزيادة في الويب وسرعة محركات البحث في التغطية.

العمري. بمعنى أنها غير متحيزة لمنطقة جغرافية أو فئة عمرية أو حتى جنس معين. وتشتمل هذه النوعية من البوابات على خمس فئات من الخدمات هي:

12.1 محرك بحث يسمح باسترجاع صفحات ومصادر المعلومات التي تم تجميعها في البوابة.

12.2 الأدلة الموضوعية التي تقوم من خلالها البوابات بعرض لمجموعة منتقاة لصفحات المعلومات في مجالات موضوعية مختلفة ومتنوعة.

12.3 خدمات التواصل وتشمل خدمات البريد الإلكتروني والدردشة والقوائم البريدية وقوائم الخدمات.

12.4 الخدمات الصحفية وتتضمن مجموعة من المواد الإخبارية التي تساعد المستفيدين من البوابة على التعرف إلى أهم التطورات في كل المجالات وفقاً لاهتماماتهم المحددة في ملف سمات المستفيدين. فإذا كان المستفيد من المهتمين بلعبة كرة القدم تبث هذه الصفحة مجموعة المواد الإخبارية الخاصة بلعبة كرة القدم أما إذا كان من المهتمين بالسياسة فتشتمل هذه الصفحة على مجموعة من الأخبار السياسية.

12.5 التجارة الإلكترونية Electronic Commerce حيث تشتمل البوابات على خدمات تسمح للمستفيد بالبحث عن السلع والخدمات التي يحتاج إليها من خلال إمكانيات التسوق الإلكتروني Electronic Shopping.

12.6 المواد المرجعية حيث تتضمن هذه المواقع إمكانيات الحصول على المعلومات المرجعية من المصادر المختلفة مثل درجات الحرارة، أسعار العملات، اتجاهات البورصات، قواميس لغوية وغيرها من المصادر التي تساعد على الإجابة عن التساؤلات السريعة والمحددة مثل: ما هي درجة الحرارة المتوقعة في مدينة نيويورك في الأيام الثلاثة التالية.

12.7 المسابقات والاستفتاءات: حيث إن هذه المواقع عادة ما تقوم بعمل مسابقات حول موضوعات معينة واستفتاءات لاستطلاع رأي المستفيدين حول موضوعات مختلفة سياسية ورياضية واقتصادية وغيرها.

ومن أمثلة البوابات العامة التي تغطي مختلف مناحي الحياة بوابة مايكروسوفت العربية http://www.arabic.arabia.msn.com/ وبوابة ياهو www.yahoo.com وبوابة جوجل www.google.com وبوابة جو www.go.com وبوابة Galaxy www.galaxy.com وبوابة http://www.excite.com/ وغيرها.

10.5.2 البوابات المتخصصة Specialized Portals

يسعى هذا النوع من البوابات إلى خدمة جمهور بعينه له سماته الخاصة سواء كانت سمات لغوية، حيث توجد بوابات بلغات معينة مثل بوابة العرب http://www.arabsgate.com/ وبوابة مكتوب العربية http://www.maktoob.com، والبوابات المتخصصة في مجالات موضوعية معينة مثل بوابة إسلام أون لاين http://www.islamonline.net/english/index.shtml بوابة الحاسب الآلي http://www.thecomputerportal.com/ البوابة الصحية http://www.thehealthportal.com. كما ظهر في الآونة الأخيرة العديد من البوابات الحكومية التي تقدم من خلالها خدمات الحكومات الإلكترونية Electronic Government Services مثل بوابة الحكومة الرقمية المصرية http://www.egypt.gov.eg/arabic بوابة الحكومة الرقمية الأمريكية FirstGov http://www.egypt.gov.eg/arabic/ بوابة حكومة دبي الذكية http://www.dubai.ae/ar/Pages/default.aspx. ويشار إلى هذه البوابات الموضوعية بمصطلح البوابات الأفقية Vertical Portals في مقابل البوابات الرأسية العامة.

يمكن الحصول على قائمة بالبوابات الموضوعية من خلال موقع البوابات الأفقية http://www.verticalportals.com.

وقد ظهر في الآونة الأخيرة نوع جديد من أدوات البحث والاسترجاع يعرف بالأعوان الذكية Intelligent Agent التي تسعى إلى توظيف تكنولوجيا الذكاء الاصطناعي لبناء نظم بحث واسترجاع خبيرة تتمكن من التعرف إلى احتياجات المستفيدين من خلال ما يقوم به من عمليات وما يصله من رسائل بريد إلكتروني

وما يقوم بفتحه من صفحات ويب. ويرى الخبراء في موقع Search Engines Watch www.searchengienewatch.com وموقع (www.w3c.org) أن هذه الأعوان الذكية تسعى إلى توظيف إمكانيات لغة التكويد الموسعة (eXtensible Mark Up Language (XML في بناء أدوات بحث دلالية Semantic Searching لكي تتوافق مع الجيل الجديد من الشبكة العنكبوتية الذي يعرف بالويب الدلالي Semantic Web.

وعلى الرغم من تنوع طرق الوصول إلى المعلومات على الشبكة العنكبوتية إلا أن 85٪ من المستفيدين من الشبكة العنكبوتية يصلون إلى المعلومات من خلال البحث في محركات البحث. وقد أوضحت دراسة التي أعدها معهد ستانفورد للدراسات الكمية أن البحث واسترجاع المعلومات يأتيان في المرتبة الثانية من حيث الخدمات المستخدمة بكثافة على شبكة الإنترنت، بينما يأتي البريد الإلكتروني في المرتبة الأولى (GVU, 2004).

وقد أشار كل من ني وابرنج في دراستهما إلى أن الإنترنت تعد اليوم مكتبة عامة هائلة تتيح العديد من الخدمات التجارية والمجانية جنباً إلى جنب. وأن أكثر الاستخدامات انتشاراً الآن على شبكة الإنترنت يتمثل في البحث عن السلع والمنتجات، والهوايات، وشركات الطيران، والمعلومات العامة والذي غالباً ما يتم من خلال محركات البحث. كما أوضحا أيضاً أن كل المستفيدين الذين تمت مقابلتهم أثناء إعداد الدراسة أكدوا أنهم نجحوا في واحدة أو أكثر من أنشطة جمع المعلومات اللازمة لاحتياجاتهم على الرغم من تنوع وتعقد الأدوات المستخدمة واختلاف تلك الاحتياجات (Nie & Erbring, 2000).

الخلاصة أن تقنيات البحث والاسترجاع على الشبكة العنكبوتية هي أدوات لا غنى عنها للوصول إلى مصادر المعلومات المتاحة على هذه الشبكة. وتعد محركات البحث من أكثر الأدوات استقراراً وتطوراً، وتوظف هذه المحركات أساليب متطورة لاسترجاع المعلومات إلى جانب أن هناك بعض الجوانب الجديدة في محركات البحث التي تجعل من استرجاع المعلومات على الشبكة العنكبوتية يختلف إلى حد ما عن نظم استرجاع المعلومات التقليدية.

المصادر

- Brin, Sergey and Page, Lawrence (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine http://infolab.stanford.edu/~backrub/google.html
- Bergman, Michael K. (August 2001). "The Deep Web: Surfacing Hidden Value". The Journal of Electronic Publishing 7 (1). http://dx.doi.org/10.3998/3336451.0007.104
- Bokor, G. (1999). Terminology Search on the World-Wide Web. Translation Journal (3), 1. Retreived from the www at 25, Febuaray, 2005.
- http://accurapid.com/journal/07search.htm
- Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. Information Processing & Management, 35, 141-180.
- Cutts, M. (2006). Ramping up on international webspam. Matt Cutts: Gadgets, Google, and SEO.
- Dupont, C., & Anderson, C. (2008). SearchWiki: make your own search. Online verfügbar unter http://googleblog. blogspot. com/2008/11/searchwiki-make-search-your-own. html.
- Gyöngyi, Zoltán and Garcia-Molina, Hector (2005), "Web spam taxonomy", Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005 http://airweb.cse.lehigh.edu/2005/gyongyi.pdf
- Gray, M. (1995). Measuring the Growth of the Web. cited by RF Morgan in 'An Internet Marketing Framework for the Web', Journal of Marketing Management, 12, 757-75.
- Gromov, Gregory (2000). History of the Internet and WWW- Part 8: Statistics. The Road and Corssroads. Feburary. Retrieved from the WWW at May,25, 2005.
- http://www.netvalley.com/intval/07262/main.htm?sdf=1
- Gulli, A., & Signorini, A. (2005). The indexable web is more than 11.5 billion pages. In Special interest tracks and posters of the 14th international conference on World Wide Web (pp. 902-903). ACM.
- Iskold, A. (2006). Watch Out Google, Vertical Search is Ramping Up!. Retrieved at, 4.

- Lancaster, F.W. (1998) Indexing and Abstracting in Theory and Practice. Champaign, Illinois: University of Illinois, Graduate School of Library and Information Science, 412 p.
- Lenssen, Phillip. Search Engines History. April, 2004. Retrieved from the WWW at May 14,2005
- http://blog-outer-court.com/history/
- Mayer, M. (2007). Universal search: The best answer is still the best answer. Google Official Blog, 5.
- Meng, W., Yu, C., & Liu, K. (2002). Building Efficient and Effective Metasearch Engines. In ACM Computing Survey, 34(1): pp. 48-89
- Mohamed, Khaled A. (2004). Merging Multiple Search Results for Meta-Search Engines. Ph.D Dissertation. University of Pittsburgh, USA, 200p
- Moulton, R., & Carattini, K. (2009). A quick word about Googlebombs. Retrieved June, 1.
- Price, G. D. (2002). Specialized Search Engine FAQs: More Questions, Answers, and Issues. Searcher, 10(9),42-46.
- SEO Consultants. Directory and Search Engines History. June 2003. Retrieved from the WWW at May 15, 2005.
- http://www.seoconsultants.com/search-engines/history/
- Teevan, Jaime; Adar, Eytan; Jones, Rosie; 7 Potts, Michael (2006). History repeats itself: repeat queries in Yahoo's logs. In Proceedings of SIGIR '06. http://doi.acm.org/10.1145/1148170.1148326
- Vaughan, L., & Thelwall, M. (2003). Scholarly Use of the Web. What are the Key Inducers of Links. Journal of the American Society for Information Science and Technology. 54(1), 29-38.
- Wall, Aaron. Search Marketing. History of Search Engines & Web History. Retrieved from the WWW at May, 16, 2005.
- http://www.search-marketing.info/search-engine-history/
- Weiss, Rick: On the web, research work proves ephemeral: Electronic archivists are playing catch-up in trying to keep documents from landing in history's dustbin (2003). The Washington Post. http://www.washingtonpost.com/ac2/wp-dyn/A8730-2003Nov23
- Yang, X. & Zhang, M. (2000). Necessary Constraints for Fusion Algorithms in Meta Search Engine Systems. In Proceedings of International Conference on Intelligent Technologies, Bangkok. (Accessed through Citeseer Search Engines): pp. 409-416

الفصل الحادي عشر

دراسات تمثيل المعرفة
والاسترجاع والفرز
في بيئة الويب

◄ 11 مقدمة

يشتمل هذا الفصل على مراجعة علمية تفصيلية للدراسات المتعلقة بتمثيل المعرفة بمحركات البحث وآليات تكشيفها وفرزها في بيئة الويب من خلال تحديد ملامح تلك البيئة والفرق بينها وبين غيرها من بيئات تمثيل المعرفة. ويركز الفصل بصفة أساسية على المنهجيات والقياسات المتبعة في دراسات الويب. وقد تم تقسيم الدراسات إلى: دراسات واقعية تعمل في البيئات التشغيلية، ودراسات معملية تتم في المختبرات وفي بيئات اصطناعية، ثم تناول الفصل آليات التكشيف وطرق دراستها. وسيتناول الفصل كل السبل الممكنة لدفع النتائج وترقيتها بمحركات البحث، إلى جانب عرض لطبيعة المشكلات التي تتناولتها الدراسات بغرض توضيح اتجاهات الإنتاج الفكري في هذا المجال إلى جانب طبيعة المناهج والأساليب المتبعة في دراسة تلك المشكلات. وتجدر الإشارة أن دراسات الويب مازالت من الدراسات الناشئة التي تسعى إلى البحث عن مناهج تتوافق معها من حيث البنية وطبيعة الاستخدام وهو ما دعى إلى ظهور مصطلحات جديدة في الإنتاج الفكري المتخصص في مناهج البحث للإشارة إلى هذه النوعية من الدراسات والقياسات التي تتوافق معها من أهمها مصطلح قياسات الويب Web Metrics.

◄ 11.1 التكشيف والفرز في بيئة الويب

WEB INDEXING AND Ranking

خلال الأعوام الأربعين الماضية مرت طرق وأساليب تكشيف واسترجاع المعلومات بمراحل متعددة وتطورت بشكل مذهل من خلال التجارب والاختبارات

التي أجريت عليها. ومع ظهور الشبكة العنكبوتية تم تطوير تلك الأساليب لكي تستخدم في تكشيف واسترجاع المعلومات من خلال محركات البحث ولكي تتوافق مع طبيعة البيئة الجديدة التي تعمل فيها هذه المحركات، حيث تم في بعض الأحيان تطويع هذه الأساليب، وفي أحيان أخرى تم توسيعها أو تغييرها بالكامل لكي تشمل طرقاً جديدة للتكشيف والاسترجاع والفرز.

يعتمد تكشيف الويب وما تحتويه من صفحات ومواقع على اختلاف أنواعها على أساليب التكشيف الآلي، حيث إنه الأسلوب الذي يتناسب مع طبيعة الويب من حيث الحجم Size والاتساع scalability، والتحديث Update المستمر لمحتواها. وتعد محركات البحث هي الأداة الوحيدة في الوقت الحالي القادرة على التعامل مع الويب بهذه المواصفات. وتختلف محركات البحث من حيث طبيعة المواد التي تنتقيها من مصادر الويب ومن حيث المصادر والأساليب التي تستخدمها في تكشيف تلك المواد، إضافة إلى أنها تختلف من حيث القدرات التي تتيحها لبحث المواد، هذا إلى جانب تنوع المصادر المكشفة نفسها. وهو ما يفسر النتائج المختلفة التي تسترجعها محركات البحث عندما يتم بحث الاستفسار نفسه في أكثر من محرك في الوقت نفسه.

كما تختلف محركات البحث من حيث الإجراءات التي تتبعها في تحديد حجم المادة المكشفة التي تتراوح ما بين التكشيف الانتقائي والتكشيف الشامل، حيث تعلن بعض المحركات صراحة أنها تكشف عدد N من الحروف أو N من الكلمات الأولى في الصفحات المكشفة، والبعض الآخر عادة ما يكون غامضاً في هذه الناحية. كما أن بعض محركات البحث تقوم أولاً ببناء مستخلص تشتقه من الصفحات المكشفة ثم تستخدم هذا المستخلص في تكشيف تلك الصفحات.

ومن أمثلة الأساليب المستخدمة في التكشيف على الويب ما يتم تطبيقه في محرك البحث EXCITE الذي يدعي استخدام أسلوب الاشتقاق الذكي للمفاهيم Intelligent Concept Extraction بالاعتماد على استخدام منهجية درجة التشابه Similarity Score Approach التي تعتمد على وزن المصطلحات. وتجدر الإشارة

إلى أن هذا الأسلوب يكتنفه كثير من الغموض نظراً لاعتماده بصفة أساسية على المصطلحات كثيرة التردد، وهو ما يمكن خداعه ببساطة من خلال استخدام أساليب خداع محركات بحث Search Engines Spamming or Persuasion التي تعتمد على التعرف إلى أساليب التحليل والتكشيف والفرز في المحركات بغرض دفع أو ترقية النتائج في محركات البحث Search Engine Optimization.

وتختلف محركات البحث في أساليب وإمكانيات فرز المخرجات والتي تعتمد على إجراءات وأساليب التكشيف المستخدمة بتلك المحركات، إضافة إلى نوع وحجم المعلومات المخزنة في ملفات البحث. ومن الطرق والأساليب المتبعة في فرز النتائج ما يلي (Big Search Engine Index, 2002; Chu & Rosenthal, 1996).

1. الفرز وفقاً لتردد المصطلحات

يعتمد هذا الأسلوب على تحديد درجة معينة لكل وثيقة تتراوح بين (صفر وواحد) وفقاً لعدد مرات ظهور مصطلحات البحث في الوثيقة. فالوثيقة التي يظهر فيها مصطلحات البحث 5 مرات أفضل من وثيقة أخرى ظهر فيها مصطلح البحث 3 مرات. وبالتالي فالوثيقة الأولى تسبق الوثيقة الثانية في الترتيب. كما يمكن دمج هذا الأسلوب مع حجم الوثيقة للتعرف على أهمية المصطلح في الوثيقة، ففي حالة وجود وثيقة مكونة من 1000 كلمة وظهر فيها مصطلح البحث عشر مرات، ووثيقة أخرى مكونة من 100 كلمة وظهر فيها مصطلح البحث 5 مرات، فبالدمج بين أسلوب تردد المصطلحات وحجم الوثيقة نجد أن الوثيقة الثانية أفضل من الوثيقة الأولى إحصائياً.

2. الفرز وفقاً لمضاهاة N من مصطلحات البحث

على سبيل المثال نفترض أن استراتيجية بحث تتكون من 7 مصطلحات جميعها كلمات بحثية (أي لم ترد في قائمة الاستبعاد). فالوثيقة التي تشتمل على كل المصطلحات الواردة في الاستفسار أفضل من وثيقة أخرى تشتمل فقط على ستة من هذه المصطلحات السبعة والتي تكون بالتبعية أفضل من وثيقة ثالثة تشتمل على 5 مصطلحات فقط وهكذا.

3. مكان ظهور المصطلح

تعتمد هذه الطريقة على تحديد موضع مصطلحات البحث في الوثيقة، فالوثيقة التي تظهر فيها مصطلحات البحث في بدايتها مثل العنوان أو رأس الوثيقة يُفترض أنها أفضل من وثيقة أخرى تظهر فيها مصطلحات البحث في وسط أو نهاية الوثيقة.

4. تقارب المصطلحات

يشير إلى الوثائق التي تكون مصطلحات البحث فيها مجاورة لبعضها البعض والتي تعد بالطبع أفضل من وثيقة أخرى تشتمل على مصطلحات البحث في مناطق متفرقة من الوثيقة.

5. استخدام الميتاداتا

تعلن بعض محركات البحث صراحة أنها تعطي أولوية خاصة للوثائق التي تشتمل على وصف مسبق باستخدام معايير الميتاداتا، بينما يعلن عدد آخر من المحركات أنه يتجاهل الميتاداتا تماماً في عمليات التكشيف والفرز.

6. عدد الروابط

وقصد به عدد الروابط التي تتضمنها الوثيقة والتي تحدد علاقتها بوثائق أخرى إلى جانب عدد الروابط التي تستخدم في وثائق أخرى مكشفة في محرك البحث للإشارة إلى هذه الوثيقة. وتجدر الإشارة إلى أن محركات البحث لا تعتمد على أسلوب واحد في فرز النتائج، ولكنها عادة ما تستخدم أكثر من أسلوب للفرز في الوقت نفسه. وعادة ما تخفي المحركات الأسلوب الذي تستخدمه في تكشيف وفرز النتائج. مع العلم أن هذه العمليات يمكن الكشف عنها من خلال الفحص الدقيق لأساليب التكشيف والفرز في محركات البحث.

وتختلف بيئة استرجاع المعلومات على الشبكة العنكبوتية عن بيئة استرجاع المعلومات التقليدية في العديد من الجوانب منها: (Huang, 2000; Rasmussen,(2003.

1. حجم المعلومات Collection Size

فعدد الصفحات والمواقع المتاحة على الشبكة العنكبوتية ضخم جداً وفي تزايد مستمر، إضافة إلى أن هناك صفحات يتم حذفها وأخرى يتم تعديلها. ومن الجدير بالذكر أن هناك جزءاً كبيراً جداً من الشبكة العنكبوتية غير مرئي Invisible Web لأدوات البحث والاسترجاع التقليدية ويحتاج إلى أدوات خاصة للتعامل معه. وتنقسم صفحات المعلومات المتاحة على الويب إلى ثلاثة أنواع أساسية هي: الصفحات الثابتة Static Pages والصفحات الديناميكية Dynamic Pages والصفحات التفاعلية Interactive Pages. والفرق بينها ببساطة أن الصفحات الثابتة لها مواقع يمكن لأي شخص الولوج إليها، بينما الصفحات الديناميكية تحتاج إلى كلمات مرور وتحديد هوية أو إجراءات بحث مثل صفحات البريد الإلكتروني وقواعد البيانات، أما الصفحات التفاعلية فتحتاج إلى إجراء أولي أو تفاعلي من جانب المستفيد حتى تظهر على الويب مثل ما يحدث عندما نقوم باستفسار محركات بحث الشبكة العنكبوتية وتظهر لنا صفحة نتائج البحث، التي تعد في هذه الحالة صفحة تفاعلية تختفي بمجرد غلق أداة التصفح.

2. التنوع Variability

يوجد تنوع كبير في الصفحات والمواقع المتاحة على الشبكة العنكبوتية من نواحٍ متعددة مثل:

- **الحجم Size**: توجد صفحات لا تتعدى بضع كلمات وصفحات يصل حجمها إلى ملايين الكلمات.

- **هيكل البناء Page Structure**: هناك طريقتان أساسيتان لبناء المواقع والصفحات هما البناء السطح Flat Structure والذي يعتمد على سرد المعلومات بشكل تتابعي مع التقليل قدر الإمكان من الروابط الفائقة Hyperlinks التي قد تتسبب في إرباك القارئ وقطع تركيزه. أما الطريقة الثانية فتعرف بالقوائم الساقطة Drop Down Menu وهي الطريقة التي تعتمد على

استخدام الروابط الفائقة بشكل مكثف، بحيث يتم قراءة ومتابعة المعلومات من خلال قوائم أساسية تنتقل إلى قوائم أخرى. ويعد هذا النمط، من أهم الملامح المميزة للويب كبيئة لاسترجاع المعلومات، إلا أنه قد يحدث إرباك للمبتدئين في التعامل مع الشبكة العنكبوتية.

- **التركيز Focus:** يعتمد أسلوب الكتابة في بناء مواقع الويب على الأسلوب الصحفي الذي يحاول تقديم أكبر قدر من المعلومات في أقل مساحة ممكنة، إضافة إلى استخدام الروابط الفائقة للحصول على المعلومات المفصلة.

- **الجودة Quality:** حيث تعد جودة المعلومات المقدمة على الشبكة العنكبوتية من القضايا الشائكة التي تحتاج إلى بحث مضنٍ وشاق من جانب المستفيد للتأكد من صحة وسلامة المعلومات التي يحصل عليها من تلك الشبكة. فمن المعروف أن المعلومات التي تنشر على الشبكة العنكبوتية لا تخضع للمراقبة أو المراجعة وهو ما جعل من الشبكة العنكبوتية تحمل الكثير من المغالطات والمعلومات السطحية. لذلك تظهر الحاجة إلى معايير لتقييم جودة المعلومات التي تقدمها مواقع الويب. وتوجد العديد من الدراسات التي تحاول وضع معايير لضبط جودة المعلومات المتاحة على الشبكة العنكبوتية بحيث يستطيع المستفيد تقييم المصادر التي يحصل منها على المعلومات (فراج، سبتمبر 2003).

- **الدقة Accuracy:** تتميز الشبكة العنكبوتية بأنها بيئة ديمقراطية للنشر لا تخضع للرقابة أو التحكم، ما أدى إلى وجود تضارب كبير بين المعلومات المتاحة من خلالها وما يقدمه غيرها من المصادر. والغريب أن البعض يعتقد أن المعلومات المتاحة على الشبكة العنكبوتية أكثر دقة من غيرها من المصادر. والحقيقة أن الويب مثلها مثل غيرها من بيئات استرجاع المعلومات تطرح ما يقدم إليها من معلومات بصرف النظر عن الوسيط. فهي لا تختلف عن بيئة النشر التقليدية حيث يوجد بها مصادر إلكترونية يتم تحكيمها وحوكمتها بآليات صارمة للتحقق من دقة وجودة المعلومات، وبها المصادر الحرة مثل

الموسوعات المفتوحة التي تعتمد على إتاحة معلومات عامة والمدونات التي لا تخضع للرقابة أو التحكيم.

- **التنوع في أنواع الوثائق Wide Variety of Document Type:** فالوثائق المتاحة من خلال الشبكة العنكبوتية تشتمل على أشكال متنوعة مثل صفحات ومواقع الويب، ملفات البي دي إف PDF، الكتب، الدوريات، الرسائل الجامعية، صور، أصوات، ملفات فيديو وغيرها من أشكال أوعية المعلومات المتاحة في شكل رقمي. هذا إضافة إلى التنوع في الأدوات المستخدمة في إعداد هذه الوثائق مثل لغات[1] HTML, XML, XSL, JAVA SCRIPT, JAVA, CSS, PDF Maker, Office Tools,..etc..

3. التكرار في الوثائق والمواقع المتاحة على الشبكة

كثير من صفحات ومواقع الويب متاحة من خلال أكثر من خادم واحد حيث نجد الصفحة نفسها متاحة بالمحتوى نفسه من خلال أكثر من موقع في البلد نفسه أو في بلدان مختلفة، مما يؤدي إلى خلط كبير عند التكشيف والاسترجاع كما يؤدي إلى ارتفاع معدلات التداخل والتكرار بين صفحات ومواقع الويب. ويعد مقياس التداخل والتكرار من المقاييس المهمة المستخدمة في قياس فعالية أدوات البحث والاسترجاع المتاحة على الشبكة العنكبوتية (Hawking; Craswell; Thistlewaite; & Harman, 1999).

4. الروابط الفائقة Hyperlinks

الوثائق المتاحة على الشبكة العنكبوتية مرتبطة ببعضها البعض من خلال شبكة واسعة من الروابط الفائقة Network Of Hyperlinks والتي تعد من أهم الملامح الخاصة التي تميز الشبكة العنكبوتية عن غيرها من بيئات استرجاع المعلومات مثل قواعد البيانات الببليوجرافية. وقد أتاحت هذه الميزة إمكانية ربط قواعد البيانات الببليوجرافية بالنصوص الكاملة للدوريات الإلكترونية وغيرها من مصادر المعلومات الإلكترونية.

---

(1) هي مجموعة لغات البرمجة الخاصة ببناء صفحات مواقع وصفحات الويب.

5. المعالجة القبلية Preprocessing

تحتاج الصفحات ومواقع الويب المتاحة من خلال الشبكة العنكبوتية إلى معالجة قبلية Preprocessing بسبب حجمها وطبيعتها الديناميكية المتغيرة، الأمر الذي يتطلب نوعية خاصة من المصادر غير المرئية لكي تعمل على متابعة تحديث عمليات التكشيف والاسترجاع. ويقصد بالمعالجة القبلية ما تقوم به الزواحف أو العناكب Spiders or Crawlers من زيارة خوادم الشبكة العنكبوتية بغرض تجميع الصفحات ومتابعة تحديثها وهو أمر من الصعب أن يتم من دون برامج خاصة للمعالجة القبلية للوثائق لاختبارها ومقارنتها بالصفحات التي تم تجميعها من قبل.

6. الاستفسارات Queries

غالباً ما يكون حجم الاستفسارات التي توجه إلى أدوات البحث على الشبكة العنكبوتية أقصر من غيرها في البيئات التقليدية. وقد أثبت العديد من الدراسات أن الاستفسارات المستخدمة على الشبكة العنكبوتية تتراوح ما بين كلمتين إلى ثلاث بمتوسط 2.4 كلمة في الاستفسار الواحد أما الاستفسارات التي تستخدم في الاسترجاع من قواعد البيانات سواء كانت ببليوغرافية أو نصية فتتراوح ما بين 12 ـ 15 مصطلح في المتوسط (Jansen; Spink; Pfaff, 2000).

7. سلوك المستفيدين User Behavior

يختلف سلوك المستفيدين في التعامل مع بيئة الويب عن سلوكهم في التعامل مع غيرها من مصادر المعلومات مثل المكتبات وقواعد وبنوك المعلومات. فالويب تتميز بأنها بيئة تفاعلية إلى جانب طبيعتها الترابطية الديناميكية التي نتجت عن استخدام النصوص الفائقة إضافة إلى طبيعتها الديمقراطية والعالمية والاجتماعية، ما أعطاها أبعاداً سياسية وثقافية واجتماعية وميزات إضافية أخرى تفوق غيرها من مصادر المعلومات التقليدية (Cothey, 2001).

11.2 ملامح الويب

توصف الشبكة العنكبوتية بأنها فضاء واسع وموزع يتضمن كماً هائلاً من مصادر المعلومات، كما توصف بأنها مكتبة عامة ضخمة. كما وصفتها جرفيث بأنها مصدر معلومات متاح كلياً لملايين من البشر في جميع أنحاء العالم على الرغم من أنها تفتقد إلى الملامح الرسمية للمكتبة والغرض والاتجاه المحددين للمكتبات اللذين يتشكلان من خلال سياسات تنمية المقتنيات وبناء المجموعات. ومع ذلك فهي بالنسبة لعدد كبير من المستفيدين أكبر وأهم مصدر معلومات إلى جانب أنها أكثر المصادر إقناعاً بالنسبة للمستفيدين (Griffiths,1999).

على الرغم من أن حجم الويب غير مؤكد ولا يمكن التعرف عليه بدقة عملياً، إلا أن هناك بعض التقديرات لعدد الأجهزة المضيفة (Hosts) وعدد صفحات المعلومات المتاحة على هذه الأجهزة المضيفة. إضافة إلى تنبؤات عدة بمعدلات نمو الشبكة العنكبوتية (انظر على سبيل المثال حيث استخدم براي البيانات المشتقة من كشاف النصوص المفتوحة Open _ Text Index لعام 1995 حيث أنتج مساحة مرئية ثلاثية الأبعاد Three – Dimensional Visualization Area للشبكة العنكبوتية يعتمد على رؤية مؤشرات للمواقع (Pointer to Sites) والحجم أو عدد الصفحات في الموقع الواحد وعدد المؤشرات التي تخرج من الموقع إلى مواقع أخرى بالتالي فهو يعتمد على ثلاثة جوانب أساسية هي (Bray 1996; Diligenti, et el., 2000) :ـ

- عدد الروابط الخارجية External Hyperlinks التي تشير إلى الموقع.
- عدد الصفحات في الموقع الواحد Number of Web Pages.
- عدد الروابط التي تشير إلى مواقع أخرى داخل الموقع Internal Hyperlinks.

يوجد العديد من الدراسات التي تناولت ملامح الويب على أساس أنها كتلة أو مجموعة وثائق Corpus، حيث قام وودروف وزملاؤه بتحليل أكثر من 206 آلاف صفحة متاحة على الويب تم تجميعها من خلال زاحف الويب بشركة انكتومي (Inktomi Web Crawler) للتعرف على الأسماء السائدة للمواقع Domain Names،

حجم الوثائق، الأكواد المستخدمة في أعداد الصفحات، عدد الروابط الفائقة وغيرها (Woodruff, et. el, 1996). كما اختبر جرفينستيت ونوش الطبيعة متعددة اللغات للويب Multi lingual باستخدام طرق تعتمد على تردد المصطلحاتWord Frequencies في اللغات المختلفة فعلى أساس تحليل قاعدة بيانات AltaVista وجد أن اللغة الإنجليزية تعد أكثر اللغات شيوعاً على الويب وأن اللغات الأخرى في تزايد مستمر (Grefenstette; & Nioche, 2000).

وقد حاولت مجموعة من الدراسات وصف الويب في إطار نظري، فعلى سبيل المثال تناول البرت وجونج وبراباسي البناء الطوبولوجي Topological Structure للويب، حيث قاموا بتحديد المعامل d على أنه أقل عدداً من الروابط URL Links التي تحتاج إليها عند الإبحار بين زوج من الوثائق. فتوصلوا إلى أن متوسط عدد الروابط يصل إلى 19 رابطاً، وهو ما فسروه بأنه قطر مساحة الدائرة التي تربط بين أي صفحتين على الويب بالاعتماد على قياس أصغر مسافة بين أي نقطتين على الشبكة العنكبوتية والتي تتمثل في الحد الأدنى من الروابط بينهما (Albert; Jeong & Barabási, 1999).

وقام برودر وزملاؤه بدراسة الويب على أنها شكل هندسي مكون من صفحات أطلقوا عليها نهايات طرفية (Nodes) وروابط فائقة Hyperlinks أطلقوا عليها أقواس الدائر arcs. وكان ناتج دراستهم رسم شكل يمثل طبيعة الوصلات التي تربط بين صفحات الويب وبعضها البعض، وقد أوضح هذا الشكل أن هناك نقاطاً مركزية Central Core وهي نقاط بها عدد هائل من الروابط بحيث تشمل الصفحات القادرة على أن تتصل ببعضها البعض من خلال الإبحار باستخدام الروابط المتاحة في هذه النقاط المركزية، وقاموا بمقارنة نتائج دراستهم مع نتائج دراسة البرت وزملائه إلا أنهم وجدوا أنه لا يوجد مسار مباشر يربط بين 75٪ من النهايات الطرفية (الصفحات) Broder, et. el, 2000. وقد استخدم كل من هيوبرمان وادميك (Huberman & Adamic, 1999) بيانات موقع أليكسا Alexa- www.alexa.co، ومحرك بحث أنفوسيك (Infoseek) WWW.Infoseek.Go.com للتعرف إلى الطبيعة الديناميكية لزيادة صفحات الويب واكتشفا أن توزيع حجم المواقع يتبع قانون القوة Power law

والذي يظهر في شكل خطي على أساس Log–log أو لوغاريتم – لوغاريتم (وهو عبارة عن رسم بياني ثنائي الأبعاد يوضح علاقة لوغاريتم بلوغاريتم آخر). كما أوضحا أيضاً أن عدد الزوار لأي موقع والروابط التي تشير إلى هذا الموقع أو تربط الموقع بمواقع أخرى تتبع أيضاً قانون القوة.

ومن الواضح أن هذه التوزيعات مفيدة جداً حيث إنها يمكن أن تساعدنا على التنبوء بطبيعة العلاقات بين الروابط الفائقة وبمعدلات الزيادة في صفحات الويب إلى جانب سلوك المستفيدين عند التعامل مع تلك الصفحات.

وقد ساعدت الطبيعة الديناميكية للويب والتي تتمثل في معدلات الزيادة والتغير والتبديل سواء في محتويات الصفحات أو أماكن وجودها إلى جانب الإلغاء والحذف المستمر للعديد من الصفحات على أن أصبحت الويب بيئة فريدة تتميز بشكل كبير عن بيئة استرجاع المعلومات التقليدية. فمعرفة الطبيعة الديناميكية للويب يتيح مؤشر قوي يساعد محركات البحث في التعرف إلى الوقت المناسب لزيارة وإعادة زيارة الخوادم Server Re-Visiting من خلال الروبوت أو غيره من أدوات التجميع لتحديث كشافاتها وقواعد بياناتها.

وتوجد مجموعة من الدراسات التي حاولت التركيز على معدلات التغيير والتعديل والتحديث في صفحات الويب، ومنها ما قام به دوجلاس وزملاؤه بتحليل معدلات الاستجابة للمحتوى الكامل لصفحات إحدى الشركات التي لها موقع على الويب من خلال استخدام طلبات تعتمد على برتوكول تحويل النصوص الفائقة HTTP، ووجدوا أن 16.5 ٪ من المصادر التي تم الوصول إليها على الأقل مرتين تم تحديثها في كل مرة تمت زيارتها (Douglis, et. el, 1997). وقام كوهلر بدراسة مدى بقاء صفحات الويب من دون حذف أو تغيير، حيث اختبر مدى البقاء ومعدلات التغير لعينة من صفحات الويب ومواقع الويب. ووجد أن حوالي 12 ٪ من مواقع الويب و20 ٪ من صفحات الويب فشلت في الاستجابة بعد ستة أشهر. وقد ازدادت إلى 18 ٪ للمواقع و32 ٪ للصفحات بعد عام واحد، وأن 96 ٪ من الصفحات أجرت تعديلات في محتواها أو شكلها خلال 6 شهور وأن 99 ٪ من المواقع أدخلت تعديلات بعد عام واحد (Koehler,

1999) . كما اختبر لورانس وزملاؤه عناوين أكثر من 100.000 مقالة متاحة في قاعدة بيانات Research Index، ووجدوا أن عدد المقالات التي لم تعد متاحة على الويب انخفض من 53 ٪ عام 1994 إلى 23 ٪ عام 1999 وأن متوسط عدد العناوين في المصادر العلمية المتاحة على الويب يتزايد بشكل كبير باستمرار إلا أنهم توقعوا أن يحدث ثبات في معدلات الزيادة مع نهاية عام 2005 (Lawrence, et. el., 2001).

أما برونجتون وسيبنكو فقد استخدما بيانات تجريبية ونماذج تحليلية Analytic Modeling لحساب الوقت المناسب لمحركات البحث، الذي يجب بعده إعادة تكشيف صفحات الويب How Often a Search Engine Should Re-index Web Pages بالاعتماد على معاملين أساسيين هما (A and B) لقياس الحداثة حيث إن A تشير إلى احتمال أن يكون محرك البحث جارياً وحديثاً لعينة مختارة عشوائياً من صفحات الويب وذلك خلال فترة زمنية معينة (B) (Brewington & Cybenko, 2000).

ويتضح من العرض السابق أن الدراسات التي ركزت على الملامح العامة للويب قد اتخذت الاتجاه الوصفي التحليلي في كثير من الأحيان والتجريبي في أحيان قليلة. وقد تمثل هذا الاتجاه في ستة أبعاد أساسية هي:

1. معدلات الزيادة في الشبكة العنكبوتية من حيث الخوادم، والمواقع، والصفحات، والمستفيدين.. إلخ.

2. متوسط عدد الروابط الفائقة المستخدمة في صفحات الويب سواء كانت روابط داخلية أو روابط خارجية وتأثير ذلك على شهرة صفحات الويب Web Page Popularity.

3. أنواع وأحجام الصفحات والمواقع المتاحة على الويب والبرامج المستخدمة في إعدادها والأكواد التي يكثر ترددها في صفحات ومواقع الويب وخصوصاً أكواد الميتا أو الأكواد الوصفية.

4. تحديد شكل الويب من خلال رسم خرائط لاتجاهات الروابط الفائقة

والمسارات التي تتخذها من حيث المواقع الجغرافية أو اللغات أو أنواع الوثائق فيما يعرف بالبناء الطوبولوجي للويب.

5. دراسة الطبيعة الديناميكية للويب والمتمثلة في معدلات الزيادة والنقصان والحذف والإضافة والتعديل وأثر ذلك في أدوات البحث والاسترجاع وبصفة خاصة محركات البحث.

6. الوصف العام للويب من حيث التوزيع اللغوي والجغرافي والشكلي والزمني والموضوعي.

11.3 قياس الثبات في محركات البحث
Measuring Search Engine Stability

لقد فرضت الطبيعة الديناميكية للويب على محركات البحث التي تتولى تكشيف صفحات ومواقع الويب أن تكون ديناميكية أيضاً عند تعاملها مع الوثائق المتاحة في تلك البيئة المتغيرة، مما يؤدي إلى نتائج غير ثابتة ومتغيرة في عمليات البحث والاسترجاع. وقد أدى هذا التغيير الديناميكى إلى ظهور مشكلة رئيسة في استرجاع المعلومات من الويب تحتاج إلى دراسات لتشخيصها وإيجاد حلول لها. وقد قام كل من سيلبرج وايتزوني بتحليل نتائج محركات البحث من خلال تكرار البحث أكثر من مرة خلال فترات زمنية معينة. ووجدا أن هناك اختلافاً كبيراً في النتائج المسترجعة أكبر بكثير مما يمكن تفسيره وفقاً للدراسات المنشورة عن تقدير معدلات الزيادة في حجم أو تغير الويب. فأشارا إلى النتائج التي تختفي في قسم ثم تظهر مرات أخرى في النتائج العشر الأولى Top 10، وأرجعا ذلك إلى تغيير في عمليات المعالجة والتكشيف لتحديد جودة النتائج المطلوبة أثناء وقت المعالجة (Selberg & Etzioni, 2000). وفي دراسة أخرى مطولة قامت بها بارا – آلان استغرقت عاماً كاملاً انقسمت لجزئين كل جزء تم في 6 شهور: الأول لاختبار مدى ثبات عناوين المواقع والثاني للمتابعة، وجدت أن هناك عناوين مواقع تظهر وتختفي باستمرار. وأرجعت ذلك إلى التحديث المستمر في محتويات تلك المواقع وعدم الثبات في سياسة موردي

الخدمات (Bar-Ilan, 1998/9). وقد قام روزيو بمتابعة يومية لمدة اثني عشر أسبوعاً لمجموعة من المواقع المتاحة من خلال محركات البحث AltaVista and Northern Light ووجد عدم ثبات في محرك البحث AltaVista مقارنة بالمحرك Northern Light. وقد اقترحت ضرورة تجميع بيانات دورية لقياس ملامح الويب ومدى الثبات في محركات البحث (Rousseau,.1998/1999) كما أعدت بارا طريقة لتقييم أداء محركات البحث ومدى الثبات في أداء تلك المحركات خلال فترة زمنية محددة من خلال قياس العناوين التي تنساها محركات البحث (بمعنى عدد العناوين التي لا تتابع مدى تحديثها).

ويتضح مما سبق أن دراسات الثبات ركزت بصفة أساسية على مدى الثبات في عناوين المواقع من خلال الخوادم التي تتيحها إلى جانب مدى الثبات في متابعة محركات البحث للتغير في عناوين تلك المواقع. هذا وإن كانت الأولى أكثر أهمية من الثانية لأنها بالطبع تؤثر في مدى ثبات محركات البحث في متابعة العناوين التي تظهر وتختفي.

11.4 قياس التغطية في محركات البحث

من المنطقي أن نعتقد أنه عند بحث الشبكة العنكبوتية فإننا نبحث في جزء معين من الشبكة وهو الجزء الذي استطاعت محركات البحث تغطيته. ويرجع ذلك إلى طبيعة الشبكة العنكبوتية التي تتميز بأنها موزعة على نطاقات جغرافية كبيرة جداً لا يمكن لأي محرك بحث مهما كانت كفاءته وسرعته أن يستطيع تجميع كل صفحات ومواقع الويب في جميع أنحاء العالم إضافة إلى النمو المذهل والسريع في حجم الشبكة العنكبوتية الذي جعل محركات البحث على الرغم مما تتميز به من أدوات تجميع متميزة عاجزة عن متابعة وتحديث صفحات الويب هذا إلى جانب عدم قدرة الزواحف على تجميع المواقع والصفحات المتاحة فى الويب غير المرئي والويب المظلم. وقد قام كل من بهارات وبرودر بتطوير طريقة لحساب التغطية في محركات البحث بدلاً من الاعتماد على القيم المحدودة التي تنشرها المحركات حول عدد الصفحات التي تغطيها قواعد البيانات. وقد وجد الباحثان أنه من بين أكبر أربعة

محركات بحث أن التغطية تتراوح ما بين 17 – 47 ٪ من الصفحات المتاحة على الشبكة العنكبوتية. (Bharat, & Broder, 1998b) كما أوضح لورانس وجيل أنه من بين أكبر ستة محركات بحث لا يوجد أي من هذه المحركات يغطي أكثر من ثلث الصفحات المتاحة للتكشيف Indexable Web وأن أقل المحركات تغطية لا يغطي أكثر من 3 ٪ من الصفحات المتاحة للتكشيف (Lawrence & Giles, 1998b). وفي دراسة أخرى أعدها لورانس وجيل أكدا أن التغطية قد انخفضت مع النمو المستمر في عدد الصفحات، وعدم قدرة محركات البحث على ملاحقة هذا النمو، حيث أوضحا أن أكبر محركات البحث من حيث التغطية لا يغطي أكثر من 16 ٪ من الصفحات القابلة للتكشيف. وقد أوضح لورانس وجيل أن هذا التناقص المستمر في حجم التغطية يرجع إلى فاعلية التكلفة وعائد التكلفة، القيود التكنولوجية التي تفرض على سعة عمليات التكشيف والاسترجاع والقيود التي تفرض على سعة الشبكة. (Lawrence & Giles 1999) وإن كنا نتفق مع كل هذه الأسباب التي طرحت فإننا نضيف أن تركيز محركات البحث ينصب على صفحات المعلومات التي تنتج وتتاح من خوادم الدول التي تنتشر فيها خدمات الاستضافة في أمريكا وأوروبا والشرق الآسيوي، نظراً لسهولة التعرف إليها، يؤدي إلى تناقص التغطية مع زيادة حجم الصفحات والمواقع التي تنشر من دول وبلغات أخرى على الشبكة العنكبوتية.

يجب أن نشير في هذا السياق إلى أن النتائج السابقة لا يمكن الاعتماد عليها نظراً للطبيعة المتغيرة، إلا أنه يوجد العديد من المواقع التي توفر بيانات أكثر حداثة عن حجم التغطية في محركات البحث مثل:

http://www.searchenginewatch.com

http://showdowns.com

وقد قام نوتيس بقياس حجم الصفحات والمواقع المتاحة على شبكة الإنترنت بالاعتماد على تقدير حجم الصفحات المكشفة في ثمانية محركات بحث عالمية (Notess, 2004). كما قام كل من هينزينجر وزملاؤه باختبار مدى تكشيف صفحة معينة في عدد من محركات البحث وذلك بالاعتماد على أسلوب الواقعة الحاسمة Critical Incident لتقييم شمول

التغطية في محركات البحث، وذلك من خلال تتبع الروابط الفائقة للصفحة للتعرف على مدى تكشيف الصفحة الرئيسة والصفحات المرتبطة بها في كل المحركات محل الدراسة (Henzinger, et, el, 1999). كما قام كل من فوغان وثيلوال بقياس التحيز في تغطية محركات البحث العالمية Search Engines Coverage Bias وذلك من خلال المقارنة بين مدى تغطية الصفحات التجارية والحكومية المتاحة على خوادم 42 دولة. وأوضحت الدراسة وجود درجة كبيرة من الاختلاف في تغطية تلك المحركات فعلى سبيل المثال وجد أن AltaVista يغطي 82 ٪ من المواقع الفرنسية، بينما يغطي فقط ٪36من المواقع المصرية. وقد أثبتت الدراسة تحيز محركات البحث للصفحات المتاحة على خوادم في الولايات المتحدة (Vaughan & Thelwall, 2004). كما اكتشف كل من موشيوتز وكاوجشي طريقة جديدة لقياس التحيز في التغطية Coverage Biasness من خلال اختبار النتائج التي يسترجعها أحد محركات البحث ومقارنتها بالنتائج التي تسترجعها مجموعة من المحركات مجتمعة (Mowshowitz, 2002). كما قام مقداد بقياس مدى تعرف محركين بحث مختلفين على حروف اللغة العربية وقدرتهما على تكشيف واسترجاع المواد العربية، بالتالي تحقيق أعلى مقاييس التغطية للمواقع العربية (Moukdad, 2002).

وقد أوضح موقع الويب http://www.searchengienswatch.com في ديسمبر من عام 2014 أن محرك البحث جوجل يعد أكبر محركات البحث من حيث التغطية ويبلغ حجم قاعدة بياناته 20 بليون صفحة. وقد بلغ حجم قواعد البيانات لعدد من محركات البحث الشهيرة مثل Altavista, Alltheweb and Yahoo ما بين بليون إلى 5 بلايين صفحة. ويرجع تفوق محرك البحث Google إلى أسلوب التكشيف الذي يستخدمه حيث يعتمد على تحليل روابط الويب (Sullivan. ,2005a, December 11)، Web Hyperlinks Analysis.

ونظراً لأن محركات البحث تعتمد في كثير من الأحيان على الروابط الفائقة للتعرف إلى الصفحات والمواقع الجديدة، فقد أوضحت سوزان فيلدمان أنه من الصعب أن يتم تكشيف صفحة ويب لا تتضمن أي روابط فائقة، كما أوضحت في دراستها أن محركات البحث تستغرق في المتوسط 57 يوماً لكي تتعرف على أي صفحة جديدة تضاف إلى الشبكة العنكبوتية (Feldman, 1999).

ويرى هينزينجر وزملاؤه أنه نظراً لعدم قدرة محركات البحث على متابعة النمو الهائل والسريع في حجم الشبكة العنكبوتية فإنه من الأفضل أن تركز تلك المحركات على جودة عملية التكشيف، فقاموا بتطوير واختبار طريقة تعتمد على السير العشوائي Random Walk بين صفحات الويب وذلك بغرض تقدير قيمة ترتيب الصفحة Page Rank Value بين صفحات الويب، كما استخدموا طريقة بهارات وبرودر لتحديد أين تم تغطية الصفحة وتكشيفها في محركات البحث التي قاموا باختبارها. وقد وجدوا أن محرك البحث Lycos يعد أفضل محركات البحث من ناحية متوسط جودة الصفحة بالاعتماد على مقياس ترتيب الصفحة (Henzinger, et al, 1999).

11.5 تقييم التكشيف والاسترجاع من الويب

عند النظر إلى الشبكة العنكبوتية كبيئة لاسترجاع المعلومات نجد أنها بيئة معقدة للغاية. ليس فقط بسبب أن مجموعة الوثائق (صفحات الويب) تتغير باستمرار ولكن أيضاً بسبب الاختلاف الواضح بين محركات البحث من حيث عدد الصفحات التي يتم تغطيتها في كل محرك على حدة، إضافة إلى أن معلومات الصلاحية الخاصة بتلك المجموعات غير متوافرة بصفة عامة، كما أنه من الصعب تقييم مثل هذه المجموعات الكبيرة للحصول على معلومات عن مدى صلاحيتها. وعلى الرغم من ذلك، فإن الزيادة الكبيرة في أعداد محركات البحث المتاحة قاد الباحثين بشكل طبيعي إلى سؤال مهم يتعلق بأي من هذه المحركات أفضل من حيث الأداء والتغطية، ما أدى إلى العديد من الأبحاث والدراسات التي تتعلق بهذا الموضوع المهم.

وقد ميز جوردون وباثاك بين نوعين من الدراسات في هذا الإطار وهما: الدراسات التجريبية «Testimonial» والدراسات الوصفية «Shootout»، على الرغم من أن العديد من الباحثين أقاموا إجاباتهم على أساس ملامح عامة وأحداث أو تجارب غير منتظمة، فهناك عدد كبير من الدراسات التي حاولت تطبيق معايير صارمة تعتمد على نماذج تجريبية في استرجاع المعلومات (Gordon & Pathak, 1999).

وتعد المراجعات العلمية من أهم المصادر التي تساعد على التعرف إلى معايير

تقييم الأداء. ومن المراجعات المبكرة التي تمت للدراسات المتعلقة بقياس أداء محركات البحث، ما قام به سشوارتز حيث حلل في مراجعته العلمية الدراسات التي حاولت قياس أداء محركات البحث خلال الفترة من 1994 حتى 1998 (Schwartz, 1998). كما أشار أوبنهيم وزملاؤه إلى الحاجة الملحة إلى مجموعة النماذج والتجارب التي تساعد على تحديد معايير لدراسة الأداء في محركات البحث (Oppenheim; Morris & McKnight, 2000). وبالطبع قادت هذه المراجعات إلى سؤال مهم هو ما هو الشكل الذي ينبغي الاعتماد عليه عند تقييم محركات البحث؟ حيث إن الطريقة التقليدية التي تستند إلى أكثر مقاييس الأداء انتشاراً وقبولاً من جانب الباحثين والتي تعتمد على مقياسي الاستدعاء والتحقيق قد تكون قاصرة بشكل كبير عن قياس أداء محركات البحث. وتستخدم التجارب الكلاسيكية التي تتم في بيئة المعمل في هذا القياس حيث يتم التحكم في كل العوامل المحيطة ببيئة النظام من حيث مجموعة الوثائق التي تكون ثابتة، الاستفسارات التي تتاح في شكل معياري موحد، الوثائق الصالحة لاستفسار بعينة ومعروفة مسبقاً. وييسر هذا التحكم والضبط المعملي عمليات حساب ومقارنة مقاييس التحقيق والاستدعاء لمجموعة من الاستفسارات عبر مجموعة من النظم المختلفة أو لنفس النظام من خلال التنويع في المعاملات الداخلية الخاصة بذلك النظام، بينما نجد أن مقاييس الأداء في البيئات أو النظم العاملة Operational Environment أكثر تعقيداً نظراً لأن مجموعة الوثائق تتغير باستمرار ومجموعة الوثائق الصالحة لأي استفسار من الصعب حسابها عملياً. فإذا كان المستفيد منخرطاً في التجربة نجد أن هناك اختلافات عدة تظهر بين المستفيدين من حيث المعرفة العامة وخبرات البحث، إضافة إلى التعقيد الشديد في حساب صلاحية النتائج المسترجعة.

وقد أشار كل من ليتون وسرفيستافا إلى أن نتائج الدراسات التي تمت في المراحل الأولى من بناء محركات البحث لا يمكن الاعتماد عليها، نظراً لأن هناك العديد من التغييرات التي طرأت على ملامح محركات البحث وإمكانياتها والأساليب التي تعتمد عليها في عمليات التكشيف والاسترجاع. وقد أوضحا أن الجانب الأكثر أهمية في دراسات محركات البحث الآن هو عملية التطوير المستمر لطرق تقييم أدوات البحث على الويب، كما يتم تقديم أو طرح أساليب جديد ومبتكرة للتقييم (Leighton & Srivastava (1999.

11.5.1 التقييم في البيئات التشغيلية الواقعية

تعد دراسة دينج ومارشيونيني من أقدم النماذج لمثل هذه التجارب التي حاولت تقييم محركات البحث في بيئتها، حيث تضمنت الدراسة مقارنة بين الملامح العامة لكل محرك بحث، إضافة إلى دراسة تجريبية استخدمت عدد محدود من الاستفسارات واختبرت ثلاثة من أشهر محركات البحث في ذلك الوقت هي Infoseek, Lycos, OpenText. واشتملت الدراسة على تقييم النتائج الصالحة في مجموعة العشرين نتيجة الأولى Top 20 لكل استفسار. وقد توصلت الدراسة إلى أنه لا يوجد محرك بحث أفضل من الآخر وأن هناك اختلافات واضحة في معالجة الاستفسارات. وقد أدهش الباحثين في هذه الدراسة انخفاض معدل التداخل والتكرار بين محركات البحث، كما استخدما كفاءة عملية التكشيف وسرعة الاستجابة كمقاييس لأداء محركات البحث (Ding & Marchionini, 1996). وفي دراسة أخرى لتومايولو وباكر اللذين حاولا استخدام عدد أكبر من الاستفسارات وصل إلى 200 استفسار لتقييم أداء خمسة محركات البحث هي: (Magellan, Point, Lycos, Infoseek, AltaVista) بالاعتماد على معدلات التحقيق للنتائج العشر الأولى كمقياس لأداء محركات البحث (Tomaiuolo & Packer, 1996). أما شو وروزينسال فقد قيما أداء ثلاثة محركات بحث بالاعتماد على أسئلة مرجعية حقيقية تم توجيهها إلى قسم المراجع. وقد اشتمل التقييم على مقاييس أداء أخرى مثل وقت الاستجابة، واختيار المخرجات، وجهد المستفيد (Chu & Rosenthal, 1996). وقد لاحظت شو الحاجة إلى مقاييس تقييم تعتمد على أحكام المستفيد على النظام، حيث اقترحت طريقة منتظمة Systematic Methodology تتضمن الاعتماد على المستفيدين الحقيقيين الذين يقومون بجمع معلومات عن ملامح المشاركين في النظام، إضافة إلى معدلات التحقيق وترتيب المستفيدين للصلاحية Relevance Ranking By Users ورضاء المستفيدين وقيمة النتائج المسترجعة ككل. وقد تم الاعتماد على هذه الطريقة في دراسة رائدة لأعضاء هيئة التدريس وطلبة الدراسات العليا وتوصلت إلى اختلافات واضحة بين محركات البحث وذلك من خلال المقارنة بين أربعة محركات بحث هي AltaVista, Infoseek, Lycos ,OpenText (Su, 1997).

واستخدم ليتون وسيرفستافا 15 استفساراً لقياس التحقيق في 5 محركات بحث هي AltaVista, Excit, HotBot ,Infoseek, Lycos. وعلى الرغم من أن قيمة محركات البحث التي قاما بتقييمها قد تكون محدودة مقارنة بما كان متاحاً وقت الدراسة، إلا أن مقاييس التقييم التي اعتمدا عليها جديرة بالاهتمام، حيث اعتمدا على مقياس التحقيق في العشرين نتيجة الأولى First 20 الذي تم تعديله ليشمل وزن Weights للترتيب ضمن النتائج العشرين الأولى. واستخدما أحكام صلاحية ثنائية Binary Relevance Judgement ضمن خمس فئات محددة (غير خطية). (Leighton & Srivastava, 1999).

ويشير كل من جوردون وباثاك إلى أنه على الرغم من التطوير المستمر في محركات البحث إلا أنه لا توجد مقاييس لتقييم الأداء تواكب هذه التطورات، ولا يمكن توقع ظهور هذه المقاييس في المستقبل القريب. وتجدر الإشارة إلى أنه مازال هناك جدل دائر حول أفضل المقاييس لتقييم أداء محركات البحث[1]، لأن نتائج دراسات استرجاع المعلومات محكومة بما توفره محركات البحث من معلومات عن التطوير وهي معلومات محدودة جداً، كما أن الخوارزميات الجديدة إذا تم توفيرها مكدسة وكبيرة مما يصعب تطبيقها. وفي دراساتهما لمحركات البحث وجد جوردون وباثاك أن فعالية استرجاع محركات البحث تعتمد بشكل كبير على وظائف المضاهاة المتاحة لكل محرك بحث أكثر من اعتمادها على قدرات صياغة الاستفسارات وإمكانيات البحث المتاحة. كما لاحظا أيضاً انخفاض معدلات التداخل والتكرار بين محركات البحث سواء كان ذلك للوثائق الصالحة أو الوثائق غير الصالحة (Gordon, & Pathak, 1999).

ويلاحظ بمعظم الدراسات الكلاسيكية في استرجاع المعلومات التي حاولت تقييم أداء نظم استرجاع المعلومات من خلال الاعتماد على مقاييس الاستدعاء والتحقيق، أن معظم هذه الدراسات ركزت بشكل أساسي على مقياس التحقيق أو الدقة في النتائج المسترجعة، إما بسبب صعوبة قياس الاستدعاء في بيئة الويب

(1) أنظر مؤتمر استرجاع النصوص TREC Web Track, 2005 (Text Retrieval Conference).

أو بسبب افتراض ساد في تلك المرحلة وهو أن التحقيق أكثر مواءمة لاحتياجات المستفيدين من الويب. والاستثناءات قليلة في هذا الإطار منها دراسة كلارك ووليت حيث استخدما 30 استفساراً وثلاثة محركات بحث لقياس الاستدعاء فعرضا لطريقة جديدة لقياس الاستدعاء في محركات البحث تعتمد على الاستدعاء المسحوب Pooled Recall والذي يتم فيه تحديد الوثائق الصالحة من المحركات الثلاثة لكل استفسار ويتم تسجيلهم في كشاف كل محرك بحث على حدة مما يؤدي إلى أن تكون مجموعة الوثائق المسترجعة من المحركات الثلاثة أساساً لقياس الاستدعاء (Clarke & Willett, 1997). وقد ساعدت هذه الطريقة أيضاً على قياس معايير أخرى شملت التغطية، نسبة الوثائق الصالحة التي يحتويها فعلياً كشاف كل محرك.

وتسعى الاتجاهات الحديثة في قياس أداء محركات البحث نحو تطبيق معايير الجودة Quality Standards. وعلى هذا الأساس ناقش كل من ليتون وسيرفيستافا قضية الطرق التي يمكن الاعتماد عليها في تقييم محركات البحث مثل استخدام عدد كافٍ من الاستفسارات لكي تعطي نتائج إحصائية يمكن الاعتماد عليها في التحليل، تجنب التحيز في اختيار الاستفسارات العشوائية في ترتيب محركات البحث، إخفاء مصدر النتائج عمن يقومون بتقييمها للتأكد من إنصاف وعدالة عملية التقييم والبعد عن التحيز تماماً(Leighton & Srivastava, 1999). وقد قاما بتقييم دراساتهما السابقة في إطار اشتمالها على هذه المبادئ أم لا. وقدم جوردون وباثاك قائمة بسبعة معايير ينبغي أن تعتمد عليها الدراسات التجريبية التي تقيم أداء محركات البحث في بيئاتها العاملة Operational Environment لكي يمكن اعتبارها دراسة دقيقة وذات دلالة وهذه المعايير هي (Gordon & Pathak, (1999:

1. مستفيدون حقيقيون.
2. استخدام استفسارات حقيقية.
3. استخدام عدد كافٍ من الباحثين.
4. دراسة معظم محركات البحث المعروفة.
5. الاعتماد على المستفيدين أنفسهم أصحاب الاستفسارات في الحكم على جودة النتائج.

6. إجراء التجربة بشكل صارم وفقاً لمقاييس الأداء المحددة.
7. إجراء الدراسة في بيئة عاملة Operational Environment

وقد ناقش هاوكنج وزملاؤه هذه القضية المتعلقة بمعايير أداء دراسات التقييم وأشاروا إلى ضرورة ترقية ورفع كفاءة الاستفسارات Query Optimization وفقاً لإمكانات كل محرك بحث، كما قدموا قائمة مراجعة بالملامح التي يجب أن تتوافر في الدراسات المستقبلية لتقييم أداء محركات البحث في البيئات العاملة. وقد اعتمدت قائمة هاوكنج وزملائه على القائمة التي أعدها جوردون وباثاك وأضافوا إليها مجموعة من الملامح التي تتعلق بطبيعة المستفيدين الذين يقومون بالقياسات والاستفسارات التي توجه لمحركات البحث (Hawking, et. el., 2001).

11.5.2 التقييم في بيئة المختبرات الاصطناعية
Evaluation In Laboratory Environment

تتمثل المشكلة الرئيسة في تقييم استرجاع المعلومات من بيئة الويب في تنوع محتوى قواعد البيانات التي تشملها محركات البحث، هذا إلى جانب أن بناء مجموعة ثابتة من صفحات الويب وجعل هذه المجموعة متاحة للباحثين يسمح بإجراء مقارنات بين محركات البحث بالاعتماد على مجموعة البيانات نفسها. على الرغم من أن هوكينج وزملاءه أشاروا إلى أن ذلك يتطلب رغبة الشركات الراعية لمحركات البحث في استخدام هذه الطرق ونتائج هذه الدراسات وبطريقة إحصائية، فإن الاعتماد على مجموعة من الصفحات الثابتة يسمح للباحثين بفصل مكونات نظام الاسترجاع أو خوارزميات تكشيف أو استرجاع محددة في المعامل من أجل قياس تأثيرها على الأداء في محركات البحث (Hawking, et. el., 2001). ويرى كل من لاندوني وبيل أن التعاون بين الباحثين في مجال استرجاع المعلومات والباحثين في مجال دراسات الويب سوف يقود بالقطع إلى وسائل فعالة لتقييم أداء محركات البحث(Landoni & Bell 2000).

وقد بدأ خلال السنوات العشر الأخيرة الاهتمام بدراسات الويب من خلال مؤتمر

استرجاع النصوص http://trec.nist.gov (Text Retrieval Conference -TREC)، وخصص لها مساراً يعرف بمسار الويب Web Track ويهدف هذا المسار إلى إجراء تجارب لبناء مجموعات اختبار Test Collections تضاهي أو تماثل بيئة الاسترجاع على الويب. ويعقد هذا المؤتمر السنوي تحت رعاية المعهد القومي للمعايير والتكنولوجيا (National Institute of Standard and Technology (NIST) بهدف تشجيع الأبحاث والدراسات في مجال استرجاع المعلومات بالاعتماد على مجموعات اختبار كبيرة تشجع عمليات التطوير في طرق التقييم، إضافة إلى تبادل أفكار الأبحاث وتطبيقاتها في مجال استرجاع المعلومات من الويب (Voorhees, 2000a).

ويحصل المشاركون في هذا المؤتمر على مجموعات الاختبار والاستفسارات وأحكام الصلاحية التي تسحب لكل الوثائق من خلال مجموعة من المتخصصين في إعداد أحكام الصلاحية من داخل المعهد القومي للمعايير والتكنولوجيا. ويعتمد الباحثون في هذا المؤتمر على معايير تقييم موحدة Standardized Evaluation Measures. ففي عام 1997 عقد أول مسار للويب (Web Track) وتم بناء مجموعة من مجموعات الاختبار مخصصة لهذا المسار. وفي المؤتمر الثامن لاسترجاع النصوص (TREC 8) تم تجهيز مجموعة اختبار حجمها 2 جيجا بايت (WT2g) من صفحات الويب وتم استخدام هذه المجموعة الصغيرة لإجراء بعض الاختبارات البسيطة لقياس الأداء في النظم

المخصصة Ad Hoc (Hawking; et. el., 2000). وفي المؤتمر التاسع تم بناء مجموعة تشتمل على 9 جيجا بايت (WT9g)، وقد ازدادت هذه إلى 100 جيجا بايت (WT100g) وتم استخدام هذه المجموعة للمهام والاختبارات الكبيرة على الويب باستخدام استفسارات تم تجميعها من الملفات الخلفية لمحركات البحث (Search Engine Log Files (Voorhees, 2000b ويتلخص الهدف الرئيس من مسار الويب في قياس أفضل الطرق التي تم استخدامها في نظم الاسترجاع التقليدية للتعرف على المناسب منها لبيئة الويب من حيث الأداء مع مجموعات الويب، وتجميع البيانات من على الويب، هذا إلى جانب تأثير هذه الطرق في المعلومات المترابطة

Linking Information. كما شهدت هذه المسارات اهتمامات خاصة مثل المقارنة بين مخرجات الترتيب البوليني Boolean –Rank Output Comparison، وقضايا تتعلق بسرعة الاسترجاع ودور الاسترجاع المتوازي Cross Retrieval مثل الاسترجاع ما بين اللغات Cross Language Retrieval.

11.6 أساليب التكشيف
Indexing Methods

بالنظر إلى حجم وسعة ومعدلات التغيير والتعديل المستمر في الشبكة العنكبوتية يكون من الطبيعي أن تسود نظم التكشيف الآلي التي تعتمد على إمكانيات الحاسبات الآلية في عمليات التكشيف والبحث. وقد وصف لينش الحاجة إلى التكشيف اليدوي والتكشيف الآلي بأنها ضرورة يفرضها تنوع احتياجات المستفيدين وتنوع مصادر الويب، حيث يرى أن مهارات التصنيف والاختيار الدقيق التي يمتلكها المكتبيون لا بد أن يكملها قدرات وإمكانيات علماء الحاسب الآلي في ميكنة عمليات التكشيف وتخزين المعلومات. كما أن الطبيعة الديمقراطية للويب تتيح لناشري الصفحات أن يقوموا بتكشيف محتويات صفحاتهم بأنفسهم من خلال وصف محتويات الصفحات داخل الصفحات نفسها باستخدام معايير الميتاداتا أو ما يعرف بما وراء البيانات (Metadata (Lynch, 1997.

فمحركات البحث عادة ما تخفي الأسلوب الذي تستخدمه في تحديد درجة التشابه Similarity Score بين الوثيقة ومصطلحات الاستفسار، ولكنها في الغالب تعتمد على طرق الوزن Weight من خلال تحديد قيمة لكل وثيقة وفقاً لخوارزميات وزن المصطلحات المعروفة Term Weighting Schemes، ثم يتم ترتيب الوثائق في النهاية وفقاً لأسلوب الوزن المستخدم. ولكن محركات البحث عادة ما تستخدم أكثر من معامل واحد لتحديد ترتيب الصفحة، فعلى سبيل المثال نجد أن محرك البحث HOTBOT يدمج أكثر من طريقة معاً لترتيب وفرز النتائج المسترجعة منها تردد المصطلحات، موضع المصطلح في الوثيقة، طول الوثيقة، وجود الميتاداتا. وتعتمد أساليب التكشيف على الويب على مجموعة من الأساليب التي سنوضحها فيما يلي:

11.6.1 التكشيف بواسطة الناشرين على الويب
Indexing By Web Publishers

يمكن للأفراد أو المؤسسات التي تضع صفحات معلومات على الشبكة العنكبوتية أن تقوم بتكشيف محتويات هذه الصفحات من خلال إتاحة مجموعة من الكلمات المفتاحية التي تصف بدقة هذه الصفحات والتي يمكن أن تستخدم عند تكشيف هذه الصفحات من خلال محركات البحث. من الناحية النظرية هذا يتيح على الأقل للأفراد والمؤسسات أسلوباً لتوجيه محركات البحث عندما تقوم بتكشيف صفحاتهم من خلال استخلاص المصطلحات الممكنة لتكشيف الصفحات. ويوجد العديد من الدراسات التي تمت على هذا الأسلوب. كما ظهر العديد من الخدمات التجارية والشركات التي تقدم العديد من الإرشادات التي تساعد الأفراد والمؤسسات على وضع المصطلحات المناسبة عند تكشيف صفحاتهم، وتعمل هذه المؤسسات بصفة خاصة على تغيير ترتيب الصفحة بحيث يمكن أن تظهر الصفحة ضمن مجموعة النتائج الأولى في البحث فيما يعرف بالترقية أو تعظيم الفائدة في محركات البحث Search Engines Optimization. بعض هذه المؤسسات تمارس أساليب غير أخلاقية لتغيير ترتيب الصفحات (Stanley, T. (1997b.

ويعتبر كود الميتا (Meta-Tag) ـ أحد أكواد لغة تكويد النصوص الفائقة (Hyper Text Markup Language (HTML ـ من أكثر الوسائل التي يمكن أن يعتمد عليها ناشرو الويب من أجل إعداد ميتاداتا تساعد على وصف المحتوى الموضوعي لتلك الصفحات، وخاصة في حقل الكلمات المفتاحية Keywords وحقل الوصف Description. وتخزن هذه المعلومات داخل الملف النصي لصفحات الويب. وتجدر الإشارة إلى أنه ليست كل محركات البحث تقوم بتكشيف أكواد الميتا Meta-Tag فعلى سبيل المثال نجد أن FAST, Google, Northern- Light على وجه الخصوص لا يقومون بتكشيف هذا الحقل نظراً لأنهم يعتبرونه حقلاً مخادعاً وغير حقيقي لأنه يعتمد على محاولة إقناع محركات البحث المعروفة بـ (AltaVista, Infoseek) Turner & Brackbill, 1998.

وقد قام كل من ترنر وبركبيل بتقييم تأثير الميتا تاج في ترتيب الصفحات لمجموعة

صغيرة من الوثائق تم إعدادها خصيصاً لهذه الدراسة، حيث اشتملت على مزيج من الأكواد. فقد اشتملت مجموعة من الصفحات على حقل الكلمات المفتاحية فقط، واشتملت مجموعة أخرى على حقل الوصف فقط، كما اشتملت مجموعة ثالثة على كل من حقل الكلمات المفتاحية وحقل الوصف معاً، بينما خلت مجموعة رابعة من الصفحات من أي من حقول الميتا تاج. وقد وجد الباحثان أن حقل الكلمات المفتاحية على وجه الخصوص ساعد بدرجة كبيرة على تحسين موقع الصفحات في كل من (AltaVista, Infoseek) Turner & Brackbill, 1998.

إلى أي مدى يعتمد ناشرو الويب على استخدام أساليب التكشيف المتاحة من أجل وضع ميتاداتا لوصف صفحاتهم؟ هذا سؤال من الصعب الإجابة عنه بصورة مباشرة نظراً لأن الدراسات التي أعدت حتى الآن تختلف عن بعضها البعض من حيث مصدر الوثائق، معالجة الميتاداتا من خلال التجميع الآلي للصفحات باستخدام برامج تحرير صفحات الويب. فبفحص أكثر من ألف صفحة ويب في بولمير للعلوم Polymer Science وجد كين وويزلي أن 24 ٪ فقط من الصفحات تضمنت واحداً أو أكثر من حقول الميتا – تاج، وعند تقييمها وجد أن المحددات Attributes يساء استخدامها بشكل واضح (Qin & Wesley, 1998). وقد لاحظ كل من لورانس وجيل ندرة استخدام حقول الميتاداتا في الصفحات والمواقع التي قاموا بفحصها حيث وجدا أن 34 ٪ من الصفحات تتضمن حقلاً مبسطاً للكلمات المفتاحية و/ أو الوصف وأن 0.03 ٪ (أقل من 1 ٪) تستخدم معيار دبلن المحوري (Lawrence & Giles, 2002). وفي عينة عشوائية مجمعة لصفحات الويب تم تجميعها من دليل البحث Yahoo وجد كرافين أن 57 ٪ من الصفحات تستخدم الميتا تاج وأن 26 ٪ من الصفحات تتضمن حقولاً للوصف، بينما استخدم 628 موقعاً معيار دبلن المحوري لوصف الصفحات (Craven, 2000).

وقد ذكرت العديد من الدراسات أن مشكلة تكشيف صفحات الويب تتمثل في قدرة ناشري الويب Web Publishers على معالجة الترتيب من خلال وضع كلمات مفتاحية مكررة في الصفحات لخداع محركات البحث، وهو ما يشار إليه بالعديد من المصطلحات مثل Search Engine Persuasion .Keyword Spam, Spam-Indexing,

Stuffing, . ونظراً لأن تردد المصطلحات من العوامل المهمة في خوارزميات الفرز والترتيب Ranking Algorithms التي تستخدمها محركات البحث فإن تكرار كلمات أو جمل معينة ـ سواء كان ذلك في حقول الميتاداتا أو في النصوص غير المرئية Invisible Text (باستخدام حروف مطبعية صغيرة بنفس اللون المستخدم في خلفيات الصفحات) لذلك تظهر هذه الكلمات في النص المصدري للصفحة ولكنها لا تظهر في الشكل المعروض على الويب، من خلال أدوات التصفح بحيث لا يمكن للعين المجردة أن تراها – يساعد على رفع ترتيب الصفحة ضمن مجموعة الصفحات المكشفة والمسترجعة. هذه الطريقة في معالجة الصفحات المكشفة تستخدم كميزة تجارية من خلال رفع منتج معين في الترتيب عن غيره من المنتجات المنافسة له في السوق أو قد يجذب مستفيداً إلى موقع معين لا يضاهي احتياجاته المعلوماتية.

وعلى الرغم أنه توجد العديد من صفحات الويب التي قد تحتاج إلى مستوى أدق من التكشيف من الذي توفره محركات البحث ولكن كل الحقائق تؤكد أن قدرة التكشيف اليدوي على أداء هذه المهمة محدودة جداً خاصة في الجزء القابل للتكشيف في الويب Indexable Web. والوضع قد يكون مختلفاً بالنسبة للجزء الخفي من الويب Hidden / Invisible Web ويقصد به مجموعة الصفحات الديناميكية والتفاعلية التي تخزن في قواعد البيانات أو يتم تجميعها حسب الطلب. وسوف نركز فيما يلي من مناقشات على أساليب التكشيف الآلي كما تؤديها محركات البحث في بيئة الويب.

11.6.2 التكشيف في محركات البحث

يوجد عدد قليل جداً من الدراسات التي تصف محركات البحث من حيث بنائها والطرق والخوارزميات التي تستخدمها في عمليات التكشيف والبحث والفرز، على الرغم من أن هناك العديد من المواقع التي تحاول وصف هذه العمليات إلا أنها مواقع لا يمكن التأكد من صحة معلوماتها نظراً لما تفرضه محركات البحث من سرية وتكتم على أساليب التكشيف والفرز التي تستخدمها. ويرجع ذلك بصفة أساسية إلى المنافسة الشرسة بين محركات البحث التي تبلغ استثماراتها الآن ملايين

الدولارات حتى إن اثنين من هذه المحركات هما Google & Yahoo يحتلان قمة معدلات الربح التي تحققها شركات تطبيقات الإنترنت في السنوات الأخيرة. وقد أشار كل من جوردون وباثاك إلى أن الخوارزميات الدقيقة التي تستخدمها محركات البحث في عملية التكشيف والاسترجاع غير معلنة وتعدها المحركات أسراراً مهمة، ولكن يمكن التعرف إليها بشكل غير مكتمل من خلال الفحص الدقيق لملفات دعم المستفيدين وملفات المساعدة والأسئلة كثيرة التردد FAQ. والاستثناء الوحيد من بين محركات البحث يتمثل في جوجل Google حيث نشر كل من برين وباج بعض التفاصيل عن الخوارزميات المستخدمة في التكشيف والفرز في محرك البحث جوجل. كما توجد بعض التفاصيل التي تم نشرها من خلال التجارب التي تم إجراؤها لبناء محركات بحث غير تجارية من أجل توفير طرق للتقييم. كما قدم العديد من الدراسات وصفاً عاماً لمكونات محركات البحث منها وصف أرسوا وزملائه بنية محركات البحث. فقد أشاروا إلى أن معظم محركات البحث تتكون من نموذج عام له بنية مركزية تتمثل في كشاف ومحرك الاسترجاع. ويشتمل أي محرك بحث على مجموعة من المكونات الرئيسة تتمثل في الزاحف أو الروبوت وهو برنامج حاسب آلي يقوم دورياً بمسح الشبكة العنكبوتية من خلال تتبع الروابط ويسترجع الصفحات من أجل تكشيفها. ثم يقوم برنامج التكشيف باستخلاص الكلمات (أو بعض أجزاء من الكلمات)، أو في بعض الحالات النصوص الفائقة Hyper Text من كل صفحة من الصفحات التي يقوم بتكشيفها ثم يقوم ببناء كشاف من هذه الكلمات المشتقة. ويتكون محرك الاسترجاع من نموذج الاستفسار Query Module الذي يتلقى الاستفسارات من المستفيدين ونموذج الفرز Ranking Module الذي يقوم بمقارنة الاستفسارات بالمعلومات المتاحة في الكشاف ثم ينتج في النهاية قائمة مرتبة بالصفحات وفقاً لعلاقتها بمصطلحات الاستفسار (Arasu, et., et., 2002). وتصميم هذه المكونات يثير سؤال بحثي مهم يرتبط بإمكانيات أداء محرك البحث بمعنى إلى أي مدى تؤثر بنية محرك البحث في أدائه من حيث التكشيف والاسترجاع. وتعد الزواحف من أهم مكونات أي محرك البحث والتي حظيت بعناية خاصة في دراسات محركات البحث.

11.6.2.1 الزواحف CRAWLERS

تتعامل الزواحف مع الشبكة العنكبوتية على أنها شكل Graph، فمن خلال استخدامها لمجموعة محددة من معينات المصادر المحددة (Uniform Resource Locator-(URLs كنقاط ارتكازية، تقوم هذه الزواحف بمسح الشبكة العنكبوتية إما على اتساعها أو عمقها بمعنى أنها إما أن تنتقل من صفحة واحدة ثم تتبع كل الصفحات المرتبطة بها من خلال تتبع الروابط الفائقة المتاحة داخل هذه الصفحة أو أن تتبع رابط فائق واحد من كل صفحة تقابلها حتى تنتهي من العمق المطلوب في تتبع الروابط والذي يتراوح ما بين 3ـ10 روابط في العمق الواحد.

وقد تناولت الدراسات موضوع الزواحف من ناحية الفعالية والكفاءة في الحصول على الصفحات بغرض التكشيف. وعلى الرغم من الارتباط الوثيق بين الفاعلية والكفاءة لأن خوارزمية الزاحف الفعال تقوم بحفظ المصادر مما يرفع من جودة قاعدة البيانات ويجعل أدوات التكشيف تؤدي عملها بكفاءة، إلا أن معظم الدراسات ركزت على الفعالية أكثر من الكفاءة. ومن القضايا التي تمت معالجتها في هذا الإطار هو كيف يمكن وضع أولويات معينة لمعين المصادر الموحد من أجل الحصول على أفضل الصفحات وذلك نظراً لمحدودية قدرة تلك الزواحف على تجميع كل الصفحات المتاحة على الشبكة العنكبوتية.

وقد قام كل من شو وزملائه بوضع نموذج لترتيب ميعنات المصادر الموحد (URLs) من حيث الأهمية يعتمد على مصفوفة تحدد أهمية الصفحات. وقد أوضحوا أن نموذج ترتيب الـ URLs الجيد يجعل من الممكن الحصول على جزء مهم جداً من الصفحات المتاحة على الشبكة العنكبوتية، بالتالي فان هذا الترتيب يساعد على الاختيار من بين الصفحات من أجل الحصول على الصفحات المهمة والتخلي عن الصفحات الأقل أهمية وهو أسلوب معروف لدى المكتبيين منذ القدم (Cho, Garcia-Molina, , & Page, 1998). وقد استخدم كل من ناجورك ووينر ترتيب الصفحة Page Rank كأساس لتحديد جودة المصفوفة ووجدوا أن استراتيجية الزحف التي تعتمد على التجميع الموسع أولاً (بمعنى الانتهاء من كل الروابط في

الصفحات المصدرية قبل الانتقال إلى الصفحات الثانوية) تعمل بكفاءة أعلى وتوفر مجموعة ذات جودة عالية من الصفحات في المراحل الأولى من عمل الزاحف مما يجعلها تتفوق على الزحف العميق (Najork & Wiener, 2001).

يعد تحديد الوقت المناسب لإعادة زيارة الصفحات Page Revisiting من المشكلات المهمة التي تتعلق بعمل زواحف محركات البحث. وقد اقترح كوفمان وليو وويبر تحليلاً نظرياً للوقت المثالي لإعادة زيارة الصفحات يعتمد على معدلات التغيير والتعديل في الصفحات (Coffman, et., el., 1998). ومن المشكلات التي تؤثر في جودة وكفاءة عمليات التحديث في قواعد البيانات ترتيب وتردد زيارة الصفحات بمعنى ما هو الترتيب الذي يجب أن يتبعه الزاحف عند إعادة زيارة الصفحات؟ وما هو عدد مرات الزيارة من أجل تحديثها؟ وقد ناقش أرسو وزملاؤه الأعمال التي تم إنجازها لتحديث الصفحات واختبارها بدقة بغرض تكشيفها في محركات البحث (Arasuet., el, 2000).

ومن القضايا الأخرى التي تمت معالجتها في هذا الإطار تخفيف العبء عن الخوادم التي تزورها الزواحف والتنسيق بين مجموعة من الزواحف في عمليات الزيارة بغرض تخفيف الحمل عن الخوادم Server Load بدلاً من زيارتها في الوقت نفسه. وقد اقترح كوفمان وليو وويبر نموذجاً خطياً Queuing Model لترتيب عمليات الزيارة. يعتمد على معدلات الإفادة من الخوادم بمعنى أن يتم تحديد ساعات الذروة في التعامل مع الخوادم وتجنب زيارتها في تلك التوقيتات حتى تتمكن من تقديم خدماتها للمستفيدين على أن تقوم الزواحف بزيادتها في غير أوقات الذروة (Coffman, et., el., 1998).

وتجدر الإشارة إلى أن معظم الزواحف تقوم بتقديم معلومات عن الصفحات من أجل تكشيفها. ويتم تخزين هذه المعلومات في مستودعات للوثائق بمحركات البحث تربط بين معلومات التكشيف وهذه الصفحات في مواقعها. ومن البدائل التي يمكن أن تساعد الزواحف في أداء هذه الوظيفة استخدام أساليب التكشيف الموزع Distributed Indexing وتخزين نسخة مخبأة من الصفحات فيما يعرف بالنتائج المخبأة Cashing of Results في نظام الحصاد Harvesting System والذي يمكن تمثيله من خلال أرشيف الويب (Bowman, et. e; 1995). ومن الجدير بالذكر أن

محرك البحث جوجل يوفر هذه الخدمة وبدأ الكثير من المحركات تتخذ المنحى نفسه في تخزين نسخ احتياطية من صفحات ومواقع الويب في الأرشيفات الإلكترونية.

وبدلاً من زحف وتجميع أجزاء معينة من الشبكة العنكبوتية يمكن للزواحف أن تركز على مجالات موضوعية معينة، حيث تسعى الزواحف إلى التركيز على هذه المجالات مما ييسر عمليات التجميع. مما يجعلها أكثر شمولاً في التغطية لهذه المجالات إضافة إلى سهولة ودقة عمليات التجميع فيما يعرف بالزواحف المتخصصة Specialized Crawlers أو الزحف المركز Clarke, et. El., 2000 (Focus Crawling). وعلى الرغم من ذلك فإن تقييم أداء الزواحف المتخصصة عملية صعبة جداً نظراً لأن الصفحات الصالحة عادة ما تكون غير معروفة. وقد اقترح اوميرا وباتل نموذجاً لبناء وصيانة كشافات متخصصة في مجالات موضوعية معينة تصلح للنظم الموزعة. O'Meara & Patel 2001.

ويرى كل من ديلجينت وزملائه أن تطبيق نموذج النظم الموزعة Distributed System Model في عمليات التكشيف يعتمد على أشكال معينة توضح مسار الزواحف الموزعة مما يعني أن الزواحف تتجه نحو التطبيق كأداة فردية في بيئة الحاسبات الشخصية بمعنى أنها يمكن أن تتعامل مباشرة مع الصفحات التي يتعامل معها جمهور الإنترنت (Diligenti, 2000). أي أنها بدلاً من تجميع الصفحات من خلال الخوادم فإنها يمكن أن تقوم بتجميع الصفحات من خلال زيارة الحاسبات الشخصية لمستخدمي الإنترنت. وتجدر الإشارة إلى أن هذا الأسلوب لا يمكن التعويل عليه كثيراً نظراً لتوجه كثير من المحركات الكبيرة إلى تطوير إمكاناتها بحيث تصبح بوابات ويب، بالتالي تحتاج إلى متابعة أكثر دقة للخوادم المتاحة على الويب لتقديم خدمات أكثر فعالية للمستخدم وفقاً لاحتياجاته الخاصة فيما يعرف بإضفاء الطابع الشخصي Personalization.

11.6.2.2 تقييم خوارزميات الفرز والترتيب
Evaluation Ranking Algorithms

تعتمد بحوث ونظم استرجاع المعلومات على عدد من الوسائل أو الأساليب في التكشيف والاسترجاع من أشهرها النموذج البوليني Boolean Model، نموذج مساحة

الزاوية Vector Space Model، والنموذج الاحتمالي Probabilistic Model. ومن النتائج الشائعة في هذه النماذج الثلاثة أسلوب جذع الكلمات Keyword Stemming، استخدام قوائم الاستبعاد Stop Lists لاستبعاد الكلمات الشائعة، استخدام نظم تردد ووزن المصطلحات Term Frequency and Term Weighting Scheme مثل نموذج tf*idf (Term Frequency) * Inverse Document Frequency بمعنى تردد المصطلحات مضروباً في عكس تردد الوثائق، إلى جانب معاملات التشابه Similarity Coefficients لحساب درجة التشابه بين مصطلحات الاستفسار ومصطلحات الوثائق (Korfhage, 1997).

ومن العيوب التي تعاني منها محركات البحث كأدوات أو نظم استرجاع معلومات ارتفاع عدد النتائج المسترجعة التي تصل إلى آلاف وأحياناً مئات الآلاف من الصفحات، وانخفاض معدلات التحقيق في تلك النتائج، وعدم قدرة تلك المحركات على الاحتفاظ ببنية النصوص الفائقة hypertext Structure للوثائق المسترجعة بمعنى الاحتفاظ بقائمة النتائج المسترجعة، وضعف تلك المحركات في معالجة استفسارات المفاهيم العامة (General Concept Queries) Kao, et. el., 2000). وقد تم استخدام الأساليب المعروفة في استرجاع المعلومات لتقييم أداء أدوات الاسترجاع في بيئة الويب في السنوات العشر الأخيرة. ثم تمت إعادة تقييم هذه الأساليب لكي تتناسب مع تلك البيئة الديناميكية كما تم اختبارها في بيئات شبيهة لبيئة الويب Web Like Environment من خلال استخدام أساليب محاكاة الويب Web Simulation في معامل ومختبرات تقييم نظم استرجاع المعلومات التي توفرها مؤتمر استرجاع النصوص[1] TREC (الذي يعقد سنوياً لتقييم أساليب الاسترجاع المتطورة (Hawking, et., el, 2000).

وقد كانت محركات البحث المبكرة تكشف فقط أجزاء من صفحات الويب ولكن مع الوقت تطور أداء تلك المحركات لتكشف النصوص الكاملة لصفحات الويب، ويمكن التماس التفاصيل الكاملة لخصائص محركات البحث من خلال مراجعة موقع www.searchengineswatch.com.

---

(1) TREC: Text Retrieval Conference

ولكن التفاصيل الكاملة عن أسلوب وزن الصفحات في الكشافات ووسائل تحديد الصلاحية من أجل الفرز تعدها محركات البحث ملكية خاصة ومعلومات سرية لا يمكن الإفصاح عنها، وعلى الرغم من ذلك توجد العديد من الدراسات التي قدمت أساليب لفرز النتائج يمكن استخدامها لمعالجة النتائج المسترجعة من محركات بحث الويب. فقد قام كل من يوونو ولي بتقييم أربع خوارزميات لفرز النتائج تعتمد على مضاهاة المصطلحات Keyword Matching والروابط الفائقة Hyper Links هذه الطرق هي (Yuwono & lee, 1996).

- تنشيط الانتشار البوليني Boolean Spreading Activation
- الأكثر استشهاداً Most cited
- نموذج تردد المصطلحات عكس تردد الوثائق القائمة على مساحة الزاوية Tf* idf Vector Space Model
- تنشيط انتشار الزاوية والتي تدمج بين نموذج مساحة الزاوية وتنشيط الانتشار Vector Spreading Activation

ومن الواضح أنه يمكن تقسيم هذه الأساليب الأربعة إلى: أساليب تعتمد على تردد المصطلحات، وأساليب تعتمد على الاستشهادات والروابط بين الصفحات. وقد توصلت الدراسة إلى أن الأساليب التي تعتمد على تردد المصطلحات تعمل بكفاءة أكبر من أساليب تحليل الروابط والاستشهادات. كما اقترحا أيضاً استخدام الاستفسارات القصيرة لأنها تعمل بشكل أكثر كفاءة من الاستفسارات الطويلة مع كل من أساليب حساب الكلمات وأساليب تحليل الروابط والاستشهادات. كما أكد كلارك وزملاؤه أن مقاييس التشابه المعياري Standard Similarity Score تعمل بكفاءة أكبر مع الاستفسارات القصيرة. وقد ساعدت نتائج هذه الدراسة على تطوير أساليب لفرز النتائج تعمل بكفاءة مع استفسارات الويب التي عاده ما تتكون من عدد قليل من الكلمات. ومن المشكلات التي عالجتها دراسات البحث والاسترجاع على الويب مشكلات حجم الكشافات وتنظيم الملفات المتعلقة بتكشيف صفحات الويب (Clarke, et., el, 2000).

ومن الأسئلة المهمة التي تم طرحها في العديد من الدراسات ما إذا كانت أساليب الاسترجاع التقليدية يمكن أن تحسن من فاعلية أداء أدوات البحث على الويب. فقد استخدم سافوي وبيكورد مجموعة من صفحات الويب حجمها 2 جيجا بايت 2-Gigabyte في مؤتمر استرجاع النصوص لتقييم كفاءة أساليب متعددة لاسترجاع المعلومات. حيث قاما بتقييم أساليب مختلفة لوزن المصطلحات منها النظام الثنائي Binary System، تردد المصطلحات، تردد المصطلحات مضروباً في عكس تردد الوثائق، تطبيع طول الوثائق Document Length Normalization، كما تم تقييم استخدام قوائم الاستبعاد وجذع مصطلحات الكشاف وتوسيع الاستفسارات. وقد تمت كل هذه القياسات لمجموعة من صفحات الويب لتقييم الأداء في بيئة تشبه بيئة الويب (Savoy & Picard, 2001).

وقد حاول هاوكينج وزملاؤه فحص الطرق المناسبة للدمج بين الملامح العامة للنظم العاملة مع التجارب المعملية للتغلب على مشكلات مقارنة أساليب استرجاع المعلومات التقليدية مع استرجاع المعلومات في بيئة محركات البحث التي تختلف إلى حد كبير عن بيئة الاسترجاع التقليدية. فقاموا بمقارنة مجموعات مؤتمر استرجاع النصوص التي تم تجميعها في المؤتمر السابع TREC - 7 من خلال استخدام هذه المجموعة في فحص كفاءة خمسة محركات بحث من خلال استخدام استفسارات قصيرة تشبه إلى حد كبير الاستفسارات التي توجه إلى محركات البحث. وتوصلت الدراسة إلى أن محركات البحث الخمسة تعمل بكفاءة أقل من متوسط كفاءة محركات البحث التي تستخدم في مؤتمر استرجاع النصوص (Hawking, et. el., 2000).

11.6.2.3 استخدام الروابط الفائقة في التكشيف
Hyperlinks For Indexing

تعد الروابط الفائقة التي تربط بين صفحات الويب من أهم الملامح التي تميز الشبكة العنكبوتية. وعادة ما ينظر إلى هذه الروابط على أنها وسائل الإبحار والتصفح الأساسية بالشبكة العنكبوتية. ومع ذلك فإن الروابط الفائقة تتضمن معلومات يمكن

استخدامها عند تكشيف واسترجاع صفحات الويب. وترجع أهمية المعلومات التي تحويها الروابط الفائقة ليس فقط إلى قيمة الروابط، ولكن أيضاً إلى أهمية الوثائق المرتبطة بالوثائق المصدرية، ومدى شعبيتها، والتي يمكن تحديدها من خلال كثرة الإشارة إلى وثيقة معينة مما يعني أهمية هذه الوثيقة وارتباطها بعدد كبير من الوثائق.

وقد طور كلينبرج نظرية الروابط الناتجة عن البحث الموضوعي Hyperlink-Induced Topic Search HITS، والتي عادة ما تعرف بنظرية النقاط الارتكازية والأسانيد Hubs and Authorites. ومن المهم التعرف إلى مفهوم النقطة الاتكازية والأسانيد في هذه النظرية.

النقطة الارتكازية Hubs: هي عبارة عن الصفحة التي تشير إلى مكان وجود المعلومات بالتالي فهي تؤشر إلى عدد كبير من الأسانيد. على سبيل المثال دليل البحث يعد نقطة ارتكازية، أو صفحة قائمة المقررات بموقع الجامعة، بالتالي فالنقاط الارتكازية تشبة قائمة المحتويات أو الكشاف.

الأسانيد Authorities: كما أن السند هو الموقع الذي توجد به المعلومات والذي يرتبط بالعديد من النقاط الارتكازية. فعلى سبيل المثال الصفحات التي تشتمل على المعلومات الواقعية مثل صفحة المقرر بموقع الجامعة أو صفحة المجلة التي يوجد بها المقالات.

وأشار إلى أنها الصفحات التي تتضمن عدداً كبيراً من الروابط التي تربطها بمجموعة من الصفحات الاستنادية الصالحة Relevant Authoritative Pages والأسانيد Authorities (وهي الصفحات التي يشار إليها من خلال عدد من النقاط الارتكازية). فالاستفسارات الواسعة التي تتضمن عدداً كبيراً من الوثائق الصالحة، عادة ما تعمل على استرجاع كل من الوثائق الصالحة وأسانيد Authoritative (أي الوثائق المرتبطة بالوثائق الصالحة، وهو ما عرف في عالم قواعد البيانات الببليوجرافية فيما بعد بالوثائق المرتبطة أو الشبيهة).

وقد اقترح كلينبرج خوارزمية النقاط الارتكازية والأسانيد لكى تستخدم في تحديد

الصفحات الاستنادية Authoritative بالاعتماد على بنية الروابط، وللتعرف على مجموعة متميزة من الوثائق الصالحة المرتبطة ببعضها البعض. وقد أحدث هذا النموذج طفرة كبيرة في محركات البحث التي طورت من أساليب التكشيف والبحث بحيث يمكن استرجاع الصفحة والصفحات الشبيهة Similar Page كذلك أصبح من الممكن استرجاع الصفحة والصفحات المرتبطة بها Kleinberg, 1998 (Related Pages).



شكل (1/11) نظرية النقاط الارتكازية والأسانيد Kao, et, el., 2000

واقترح كل من لمبل وموران طريقه أخرى للتعرف على الروابط بين صفحات الويب تعتمد على بنية الروابط Link Structure تعرف بالمشي العشوائي في الأشكال، وذلك من خلال رسم شكل لطبيعة العلاقة بين الصفحات واختيار الصفحات عشوائياً، وهي طريقة أكثر كفاءة من الناحية الحسابية من خوارزمية كلينبرج، نظراً لأنها لا تحتاج إلى كثير من المعالجات (Lempel & Moran, 2000) ولعل أكثر الطرق المعروفة والمعلنة لفرز الصفحات باستخدام الروابط الفائقة تعرف بخوارزمية فرز الصفحة PagerRank Algorithm، التي طورها باج وزملاؤه (Page et al, 1998) والتي تعمل على حساب قيمة لكل صفحة من الصفحات المسترجعة والتي تتحدد على أساس عدد الروابط في كل صفحة (من وإلى كل صفحة). وتعد خوارزمية فرز الصفحة من أهم الملامح المميزة لمحرك البحث جوجل (Brin & Page, 1998).

ولقد تم توسيع خوارزمية كلينبرج لتتضمن تكشيف النصوص إلى جانب تكشيف الروابط واستخدامها في فرز النتائج، من خلال تطوير مجمع إلى للمصادرِAutomatic Resource ARC ـ (Compiler) لكي يقوم بتجميع قوائم بمصادر الويب في موضوعات عريضة. كما ناقش كل من بهارات وهينزينجر بعض المشكلات التي تتعلق بخوارزمية كلينبرج الرئيسة والتي تشمل جرف أو سحب الموضوعات Topic Drift والتي لا تمثل موضوعات رئيسة بالنسبة للنقاط الارتكازية والأسانيد المرتبطة بها (Bharat & Henzinger,1998).

ومن الاستخدامات الأخرى للروابط الفائقة تطبيق خوارزمية تعرف بسلسلة التنشيط الواسع (Constrained Spreading Activation) بغرض توسيع نطاق البحث لتحسين معدلات الاستدعاء، حيث تبدأ هذه الطريقة بصفحة أو مجموعة صفحات صالحة Relevant Pages ثم تنتشر من خلال شبكة الروابط بين الصفحات لتقوم بحساب درجه التشابه Similarity Score لكل صفحة، ثم تحدد إلى أي درجة يمكن فرز هذه الصفحة وعرضها للمستفيد. وعادة ما تحدد المحركات نقطة معينة عندها يتم تجاهل الصفحة تماماً والنظر إلى غيرها. وقد تم تطويره ويعرف بـ (Web Search By Constrained Spreading Activation (WebSCSA لكي يعمل مرتبطاً بمحركات البحث في مختبرات TREC. وقد أثبتت نتائج الدراسات تحسين معدلات الاستدعاء باستخدام هذه الطريقة بنسبة 30٪ (Crestan & Lee, 2000).

ويرى كاو وزملاؤه إمكانية استخدام المعلومات المتاحة في الروابط الفائقة بطريقة مختلفة تعتمد على دعم تكشيف نقاط المرتكز Anchor Point Indexing وقاموا بتعريف النقاط المرتكزة على أنها مجموعة صغيرة من الصفحات المفتاحية والتي يمكن من خلالها الوصول إلى مجموعة مطابقة من الصفحات بسهولة وبسرعة مما يحافظ على بنية الوثائق المرتبطة Hyperlinked Documents على الويب، وهي تشبه النقاط الارتكازية Hubs (Kao, et, el., 2000).

وقد أشار كل من سينجال وكيسزكيل إلى أن نتائج دراسات مسار الويب في مؤتمر استرجاع النصوص أظهرت أن الاعتماد على طرق دعم الروابط فقط لا تقدم أي ميزة إضافية عن طرق تكشيف الكلمات وحدها (Singhal, & Kaszkiel, 2001). هذه النتائج تتعارض تماماً مع ما هو معروف في مجتمع استرجاع المعلومات على الويب. ومن الأسباب التي أدت بهم إلى هذه النتيجة أن بيئة مسار الويب في مؤتمر استرجاع النصوص تفضل استخدام تكشيف الكلمات المفتاحية عن تكشيف الروابط نظراً لاشتمالها على صفحات قديمة (Dated Test Collection) بمعايير الويب إضافة إلى أحكام الصلاحية التي تفضل الصفحات عن المواقع. وقد أوضحوا أن محركات البحث التي تعمل في بيئة الويب أكثر كفاءة من محركات البحث المستخدمة في TREC في عمليات الحصول على صفحة معينة لمؤسسة أو لفرد. ومع ذلك فإن كرسويل وزملاءه أشاروا إلى أن طرق الاسترجاع التي تعتمد على تحليل النصوص المرتكزة Anchor Text المشتقة من الصفحة المصدرية أو الرابط المصدري أفضل بكثير من تكشيف المحتوى النصي للصفحة الاستنادية (المرتبطة) (Craswell., Hawking & Robertson, (2001.

12.6.2.4 نموذج تحليل الروابط
Link Analysis Model

يعرف هذا النموذج في الإنتاج الفكري المتخصص بنموذج ترتيب الصفحة Page Rank. وقد ابتكر هذا النموذج طالبان من طلبة الدراسات العليا في كلية الحاسبات والمعلومات بجامعة ستانفورد وهما Sergey Brin and Lawrence Page. ويعتمد هذا

النموذج على استخدام نموذج تحليل الاستشهادات المرجعية، والذي يفترض وجود علاقة بين المقالات المستشهدة والمقالات المستشهد بها. بالتالي يمكن استخدام الاستشهاد المرجعي في التعرف إلى تأثير المقالة في المجال المعرفي بأكمله. وقد ابتكر العالم Eugene Garfield مقياساً يعرف بمعامل التأثير The Impact Factor والذي يمكن من خلاله قياس مدى تأثير دورية علمية معينة في أحد المجالات. ومعامل التأثير هو عبارة عن متوسط عدد الاستشهادات بمقالات دورية معينة خلال عام معين وذلك بعد نشرها بعامين على الأقل. ويعرف هذا المعامل أحياناً بمعامل توقيع الذكاء The Signature of Intelligence.

وكما هو الحال في العلاقة بين مقالات الدوريات والاستشهادات نجد أن روابط الويب Web Links عبارة عن صلة ديناميكية تشير إلى روابط أخرى وهذه الروابط تشير أيضاً إليها. بالتالي نجد أن نموذج ترتيب الصفحة يستخدم العلاقات القائمة بين صفحات المعلومات المتمثلة في الروابط التي تربط بين تلك الصفحات على اعتبار أنها أكثر موضوعية من غيرها من المقاييس التي تعتمد على مقاييس بشرية ذاتية. فتكرار الإشارة إلى صفحة معينة يشير إلى قيمة هذه الصفحة كما يؤكد علاقتها القوية بالعديد من الصفحات، كما أنه يعتبر من المقاييس القوية التي تشير إلى كفاءة الصفحة وجودتها وذلك مقياس في غاية الأهمية نظراً لما تعانيه الشبكة العنكبوتية وخاصة صفحات المعلومات من النقص الشديد في معايير الجودة Quality Control، بالتالي فهذا النموذج يوفر مقياساً موضوعياً لجودة الصفحات. كما يعتمد نموذج ترتيب الصفحة على استخدام طبيعة الويب المكونة من مجموعة من الصفحات المرتبطة ببعضها البعض في تحديد ترتيب وأهمية الصفحة ضمن مجموعة الصفحات المرتبطة بها (Meghabghab, 2001).

ويتم تحديد ترتيب الصفحة Page Ranking وفقاً لعدد الروابط الموجودة في الصفحة In-degree of Links والتي أشار إليها كلينبرج بالنقاط الارتكازية، وعدد الروابط التي تشير إلى الصفحة Out-degree of Links والتي أشار إليها بالأسانيد.

وقد اعتمد القائمون على بناء محرك البحث جوجل على مجموعة من الخرائط Maps التي قاموا بتجهيزها وتضمنت ما يقرب من 518 مليون وحدة من الروابط الفائقة Hyperlinks لكي تمثل عينة متميزة للعلاقات التي تربط بين صفحات المعلومات على الشبكة العنكبوتية. وقد أتاحت هذه الخرائط إجراء حسابات سريعة للتعرف إلى مدى قوة العلاقة التي تربط بين مجموعة من الصفحات، ثم ترتيب هذه الصفحات من خلال الاعتماد على تحليل ما تحويه من روابط داخلية تربطها بصفحات أخرى والروابط الخارجية التي تربط الصفحات الأخرى بها. ويتميز هذا المقياس بأنه مقياس ديمقراطي إلى حد كبير، حيث يحدد مكانة الصفحة بين غيرها من الصفحات بناء على مدى أهميتها بالنسبة للصفحات الأخرى سواء بالإشارة إلى هذه الصفحات أو بالإشارات التي تتلقاها الصفحة من الصفحات الأخرى. ويتم حساب عدد الروابط الموجودة في الصفحة وتشير إلى صفحات أخرى كما يتم حساب عدد الروابط التي تشير إلى الصفحة المصدرية ثم يتم تطبيع Normalization هذه الحسابات لتحديد قيمة تشابه Similarity Score بين الصفحة والصفحات أخرى. وتتم عملية التطبيع وفقاً للمعادلة التالية:

نفترض أن الصفحة A مرتبطة بصفحات أخرى تشير إليها (Point to it) وعددها Tn------T1 والمعامل d هو معامل ثابت ما بين (0 ـ 1) وعادة ما يأخذ القيمة 0.85 إلا في حالات استثنائية سنوضحها فيما بعد. وتشير C إلى عدد الروابط الخارجة من الصفحة وتشير إلى صفحات أخرى (Point to other Pages) بالتالي يكون حساب ترتيب الصفحة PR (A) كما يلي:

PR• (A) = (1-d) + d (PR(T1) / C (T1) +------------ PR(Tn) / C (Tn)

نلاحظ من المعادلة أن ترتيب الصفحة Page Rank يمثل توزيع احتمالي Probability Distribution لكل صفحات الويب Over Web Pages مما يسمح بترتيب الصفحات تنازلياً وفقاً لقيمة A.

ويتم حساب معامل آخر لترتيب الصفحة يعتمد أيضاً على بنية الروابط Link Structure وهو معامل يتعلق بسلوك المستفيدين عند التعامل مع الصفحة. وهذا

المعامل يتعلق بمعدلات الإفادة من صفحة معينة، مما يعني أن المستفيد يمكن أن يغير من ترتيب الصفحات وفقاً لمدى استخدامه لهذه الصفحات. ويتم تحديد مدى الإفادة من صفحة معينة وفقاً لعدد مرات النقر على الرابط الفائق المتعلق بهذه الصفحة في كل مرة تظهر فيه هذه الصفحة ضمن نتائج البحث، حيث يتم تعديل قيمة المعامل d. فإذا قام المستفيد بفتح الصفحة التي تظهر في ترتيب 3 مثلاً ولم يفتح الصفحة التي تظهر في الترتيب 1 يعتبر محرك البحث جوجل أن هذا إعلان من المستفيد أن الصفحة 3 أفضل من الصفحة 1 بالنسبة لهذا الاستفسار، مما يجعل محرك البحث يعدل من قيمة المعامل d الخاص بترتيب الصفحة 3. ومع تكرار هذه العملية من جانب أكثر من مستفيد قد يؤدي ذلك إلى ظهور الصفحة 3 قبل الصفحتين 1، 2 إذا كان سلوك كل أو معظم المستفيدين منها يسير في الاتجاه نفسه. ويعتبر هذا المقياس أيضاً من المقاييس الديمقراطية التي تميز محرك البحث جوجل عن غيره من المحركات. وتعرف عملية تعديل قيمة المعامل d برد فعل الصلاحية Relevance Feedback والذي يتوقف على مجموع سلوك المستفيدين من صفحة معينة خلال فترة زمنية معينة (Wall, 2005).

◄ 11.6.2.5 نصوص الزاوية
Anchor Text

تتم معاملة النصوص التي تعبر عن الروابط في الملف المصدري Source File ـ وهو الملف الذي يشتمل على أكواد لغة تكويد النصوص الفائقة HTML ـ بطريقة خاصة في محرك البحث جوجل. حيث تتعامل معظم محركات البحث التي تستخدم أسلوب تحليل الروابط Link Analysis مع الروابط التي توجد داخل الصفحة وتكشف النصوص التي توجد داخل هذه الروابط، بينما يكشف محرك البحث جوجل الروابط التي تشير إلى الصفحة Point to it. ولهذه الطريقة العديد من المزايا ومنها (Smith, 2005):

• PR اختصار لترتيب الصفحة Page Rank

- **أولاً**: نصوص الزاوية Anchor Text التي عادة ما تتضمن وصفاً دقيقاً لصفحة الويب يفوق ما تقدمه الصفحة في جسمها الرئيس من كلمات مفتاحية تصف الموضوع الذي تتناوله، وهو ما أثبتته العديد من الدراسات، حيث إن هذه النصوص تمثل عناوين الموضوعات الرئيسة التي تتناولها هذه الصفحات.

- **ثانياً**: نصوص الزاوية تساعد على تكشيف الصفحات التي لا يمكن تكشيفها من خلال محركات بحث نصية Text Based Search Engines بالتالي يمكن استخدام هذه النصوص في تكشيف الوسائط المتعددة Multimedia مثل ملفات الصوت، والفيديو، والصور، وبرامج الكمبيوتر، والخرائط، وقواعد البيانات.. الخ.

- **ثالثاً**: تساعد نصوص الزاوية على تكشيف صفحات لم تقم الزواحف Crawlers بتجميعها أو زيارتها، بالتالي يمكن من خلال هذا الأسلوب تجميع أكبر عدد ممكن من الصفحات أو التعرف إليها دون الحاجة إلى زيارة الخوادم التي تستضيفها، خاصة إذا ما عرفنا أن هذه الزواحف عادة ما تكون متحيزة جغرافياً ولغوياً في تغطيتها. وهو ما جعل محرك البحث جوجل من أكبر محركات البحث وأشملها من حيث حدود التغطية سواء الجغرافية أو اللغوية أو الموضوعية أو وفقاً للأسماء السائدة Domain Names. وتجدر الإشارة هنا إلى أن هذه الميزة قد تنقلب إلى عيب كبير وتسبب مشكلات كثيرة، حيث إن محرك البحث يمكن أن يسترجع نتائج لصفحات لم يزرها الزاحف مطلقاً ويتأكد من وجودها، وهنا يظهر دور المعامل d والذي يأخذ القيمة صفر في حالة الروابط الميتة Dead Links أو الروابط التي تشير إلى صفحات غير موجودة.

وقد استخدمت فكرة توسيع التغطية من خلال التعامل مع نصوص أقواس الزاوية Anchor Text Propagating للصفحات التي تشير إلى الصفحات المصدرية في محرك البحث WWW WORM وهو أول محرك بحث يتضمن زاحفاً – تم بناؤه عام 1994 – لتكشيف الصفحات غير النصية Non Textual Pages. ويعد استخدام

نصوص أقواس الزاوية عملية في غاية الصعوبة نظراً لضخامة حجم البيانات التي يتم معالجتها، حيث إن معالجة 24 مليون صفحة مثلاً تتطلب على الأقل معالجة 259 مليون نص زاوية وفقاً لما أعلنه محرك البحث جوجل في عام 2010 بمتوسط 10.8 نصوص زاوية للصفحة الواحدة (Sullivan, 2002).

وإضافة إلى استخدام الروابط ونصوص الزاوية في تكشيف الصفحات يقوم محرك البحث جوجل بتحديد موقع الرابط Link Location لتحديد أهمية الرابط في الصفحة. فتعد الروابط التي تأتي في عناوين منفصلة أكثر أهمية من الروابط التي ترد ضمن نص ما، والروابط التي ترد في المحتويات والفئات التي تتضمنها الصفحة أكثر أهمية من الروابط التي ترد في عناوين فرعية. كما يستخدم محرك البحث جوجل أساليب التكشيف التقليدية مثل أسلوب تردد المصطلحات Term Frequency، التكشيف التجاوري Proximity Indexing، وأساليب وزن المصطلحات Term Weighting Schemes.

من ثم فإن نظام ترتيب الصفحة Page Rank يعتمد على الطبيعة الديمقراطية الفريدة في الويب، وذلك باستعمال الارتباطات Hyperlinks كدليل إلى أهمية صفحة معينة. بمعنى أن جوجل يفسر الارتباط من صفحة A إلى الصفحة B على أنه تصويت من الصفحة A لمصلحة الصفحة B. لكنه لا ينظر فقط إلى كمية الأصوات (أي الارتباطات الموجهة إلى صفحة معينة)، بل يُحلل الصفحة التي تقوم بالتصويت. فإذا كانت الصفحات التي تصوِّت «مهمة»، أعطاها ذلك وزناً أكبر، وجعل الصفحات الأخرى التي تصوّت لها مهمة أيضاً.

تحصل المواقع المهمة عالية الجودة على ترتيب Page Rank أعلى، الأمر الذي يتذكره جوجل في كل مرة يُجري بحثاً. طبعاً، لا تعني الصفحات المهمة لك شيئاً إن كانت لا تطابق بحثك. لذلك يجمع جوجل بين Page Rank وتقنيات مطابقة النص Text Matching المعقدة ليجد صفحات مهمة وتلائم موضوع البحث على السواء. ولا يتوقف جوجل عند عدد المرات التي تظهر فيها عبارة معينة في الصفحة، بل يفحص كل أوجه محتويات الصفحة (ومحتويات الصفحات المرتبطة بها) ليعرف ما إذا كانت مطابقة للبحث أم لا (Google, 2005).

خاتمة

تناول هذا الفصل عرضاً للأساليب والتقنيات المستخدمة في تكشيف، وتحليل، واسترجاع، وفرز صفحات الويب من خلال محركات البحث التي تعد أهم أدوات البحث عن المعلومات على الويب. كما استعرض أساليب تقييم محركات البحث ومعايير تقييم الأداء التي اعتمدت مبدئياً على الأساليب التقليدية المعروفة في نظم استرجاع المعلومات، ثم ابتكر الباحثون مجموعة من الأساليب الجديدة التي تتناسب مع بيئة الويب وما تتميز به من طبيعة ديمقراطية وديناميكية وتفاعلية.

وقد ثبت من خلال دراسات استرجاع المعلومات أن دراسات الويب من القطاعات النشطة في الوقت الحالي في مجالات البحث والتطوير؛ نظراً لأهمية هذه البيئة للباحثين والمؤسسات المسؤولة عن التطوير على حد سواء. وقد ثبت أيضاً أن البحوث ركزت خلال السنوات العشر الأخيرة، والتي شهدت نمو وتطور محركات بحث الشبكة العنكبوتية، على ظهور ونمو أساليب مبتكرة للتكشيف والاسترجاع كان على رأسها استخدام الروابط الفائقة في تحديد شهرة صفحات الويب. كما شهدت أيضاً دوراً ملموساً لكل من معايير الميتاداتا وتحديد الفئات Categorization واستخلاص الوثائق Document Summarization وتجميع النتائج المسترجعة في عناقيد Result Clustering واستخدام الأشكال في عرض النتائج Results Visualization. هذا إضافة إلى النمو السريع والهائل في بناء أدوات بحث واسترجاع الوسائط المتعددة. وكل هذه الأساليب تسعى إلى تجميع صفحات ومواقع الويب في فئات موضوعية لتيسير التعامل معها كبيئة لاسترجاع المعلومات. وهو ما يؤكد ويبرز الدور الذي يمكن أن تلعبه أدوات أخرى لاسترجاع المعلومات مثل محركات البحث المتعددة (ما وراء المحركات) وبوابات الويب، والأعوان الذكية. كل هذه التطورات تؤكد أهمية الدور الذي تلعبه بحوث التطوير في مجال استرجاع المعلومات وأساليب التكشيف ودفع النتائج على الويب.

المصادر

- فراج، عبد الرحمن (سبتمبر 2003) تقييم مصادر المعلومات المتاحة على الإنترنت. أحوال المعرفة. س8، ع30، ص ص 70ـ66.
- Albert, R., Jeong, H., & Barabási, A-L. (1999). Diameter of the World-Wide Web. Nature, 401 (6749), 130-131
- Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2000). Searching the Web. Stanford University Technical Report 2000-37. [Online] Available at http://dbpubs.stanford.edu/pub/2000-37
- Bar-Ilan, J. (1998/9). Search engine results over time: a case study on search engine stability. Cybermetrics, 2/3, Issue 1, Paper 1. [On-line]. Available at http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html
- Bharat, K., & Broder, A. (1998b, April 24). Measuring the Web. [On-line]. Available: http://www.research.compaq.com/SRC/whatsnew/sem.htm
- Bharat, K., & Henzinger, M.R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In: W.B. Croft, A. Moffat, C.J. van Rijsbergen, R Wilkinson & J Zobel, (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98) (pp. 104-111). New York: ACM.
- Bowman, C.M., Danzig, P.B., Hardy, D.R, Manber, U., Schwartz, M.F. (1995). The Harvest information discovery and access system. Computer Networks and ISDN Systems 28, 119-125.
- Bray, T. (1996). Measuring the Web. In Fifth International World Wide Web Conference (WWW5). [Available On-line]. http://www5conf.inria.fr/fich_html/papers/P9/Overview.html
- (Also published as a special a issue at Computer Networks and ISDN Systems, Volume 28 (7-11), 993-1005.
- Brewington, B.E., & Cybenko, G. (2000). How dynamic is the Web? Computer Networks, 33, 257-276
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the Seventh International World-Wide Web Conference (WWW7), published as Computer Networks and ISDN Systems, 30, 107-117. (Longer version available at http://decweb.ethz.ch/WWW7/1921/com1921.htm).

- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stat, R., Tomkins, A., & Wiener, J. (2000). Graph structure in the web. Computer Networks, 33, 309-320. [On-line]. Also available at http://www.almaden.ibm.com/cs/k53/www9.final/
- Big Search Engine Index (2002) Available Online: September, 5, 2002 R- http://www.search-engine-index.co.uk
- Craven, T.C. (2000). Features of DESCRIPTION META tags in public home pages. Journal of Information Science, 26, 303-311
- Cho, J., Garcia-Molina, H., & Page, L. (1998). Efficient crawling through URL ordering. In: Proceedings of the Seventh International World-Wide Web Conference (WWW7), published as Computer Networks and ISDN Systems, 30(1-7), 161-172
- Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology.
- Clarke, S.J., & Willett, P. (1997). Estimating the recall performance of Web search engines. Aslib Proceedings, 49(7), 184-189
- Clarke, C.L.A., Cormack, G.V., & Tudhope, E.A. (2000). Relevance ranking for one to three term queries. Information Processing & Management, 36, 291-311.
- Clarke, C.L.A., Cormack, G.V., & Tudhope, E.A. (2000). Relevance ranking for one to three term queries. Information Processing & Management, 36, 291-311.
- Coffman, E.G., Jr., Liu, Z., & Weber, R.R. (1998). Optimal robot scheduling for Web search engines. Journal of Scheduling 1(1), 15-29
- Cothey, V. (2001) A Longitudinal Study of World Wide Web Users' Information-Searching Behavior. Journal of the American Society for Information Science and Technology. 53(2): pp. 67-78
- Craswell, N., Hawking, D. & Robertson, S. (2001). Effective site finding using link anchor information. In: Croft, W.B., Harper, D.J., Kraft, D.H. & Zobel, J. (Eds.) SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (Pp. 250-257). New York: ACM.
- Crestani, F., & Lee, P.L. (2000). Searching the web by constrained spreading activation. Information Processing & Management, 36, 585-605
- Diligenti, M., Coetzee, F.M., Lawrence, S., Giles, C.L., & Gori, M. (2000). Focused crawling using context graphs. In: Proceedings of the 26th VLDB Conference. (pp. 527-535

- Ding, W., & Marchionini, G. (1996). A comparative study of Web search service performance. In S. Hardin (Ed.), Proceedings of the 59th Annual Meeting of the American Society for Information Science (pp. 136-142). Medford, NJ: American Society for Information Science.
- Douglis, F., Feldmann, A., Krishnamurthy, B., & Mogul, J. (1997). Rate of change and other metrics: a live study of the World Wide Web. In Proceedings of the USENIX Symposium on Internet Technologies and Systems [On-line]. Available at http://www.usenix.org/publications/library/proceedings/usits97/douglis_rate.html
- Ester, M., Groß, M., & Kriegel, H. P. (2001). Focused Web crawling: A generic framework for specifying the user interest and for adaptive crawling strategies. In Proceedings of 27th International Conference on Very Large Data Bases (pp. 321-329).
- Feldman, Susan (1999). New Study of Search Engines Coverage, Information Today, vol. 16 no. 829. Retrieved from the web at April, 30, 2005. http://www.infotoday.com/newsbreaks/nb0712-1.htm
- Googel. Why We Need to Use Google. Retrieved from the WWW at August, 25, 2005 Available at http://www.google.com/intl/ar/why_use.html
- Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. Information Processing & Management, 35, 141-180.
- Grefenstette, G., & Nioche, J. (2000). Estimation of English and non-English language use on the WWW. In: Proceedings of the RIAO'2000 Conference. Paris: C.I.D. [On-line]. Available at: http://133.23.229.11/~ysuzuki/Proceedingsall/RIAO2000/Wednesday/20plenary2.pdf
- Griffiths, J.-M. (1999). Why the Web is not a library. FID Review 1(1), 229-246
- Hawking, D., Craswell, N., Bailey, P. & Griffiths, K. (2001). Measuring search engine quality. Information Retrieval, 4, 33-59
- Hawking, D., Craswell, N., Thistlewaite, P., & Harman, D. (1999). Results and challenges in Web search evaluation. In Proceedings of the 8th International World Wide Web Conference (WWW8) [On-line]. Available at http://www8.org/w8-papers/2c-search-discover/results/results.html
- Hawking, D., Voorhees, E., Craswell, N., and Bailey, P. (2000). Overview of the TREC-8 Web track. In E.M. Voorhees, and D. Harman (Eds.), Proceedings of the Eighth Text REtrieval Conference (TREC-8). (NIST Special Publication 500-246). [On-line]. Available at

- Hawking, D., Voorhees, E., Craswell, N., and Bailey, P. (2000). Overview of the TREC-8 Web track. In E.M. Voorhees, and D. Harman (Eds.), Proceedings of the Eighth Text REtrieval Conference (TREC-8). (NIST Special Publication 500-246). [On-line]
- Henzinger, M.R., Heydon, A., Mitzenmacher, M., & Najork, M. (1999). Measuring index quality using random walks on the Web. In Proceedings of the 8th International World Wide Web Conference. [On-line]. Available online at http://www8.org/w8-papers/2c-search-discover/measuring/measuring.html-
- Huang, L. (2000). A survey on Web information retrieval technologies. RPE Report. [On-line]. Available at http://www.ecsl.cs.sunysb.edu/tr/rpe8.ps.Z
- Huberman, B.A., & Adamic, L.A. (1999). Growth dynamics of the World-Wide Web. Nature, 401 (6749), 131.
- Jansen, B., Spink, A., Pfaff, A. (2000). Linguistic Aspects of Web Queries. Proceeding of the 36rd. ASIS Annual Meeting , Volume, 37: 169- 176
- Kao, B., Lee, J., Ng, C., & Cheung, D. (2000). Anchor point indexing in Web document retrieval. IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, 30, 364-373
- Kleinberg, J.M. (1998). Authoritative sources in a hyperlinked environment. Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 668-677. (A full version of the paper is available at http://www.cs.cornell.edu/home/kleinber/).
- Koehler, W. (1999). An analysis of Web page and Web site constancy and permanence. Journal of the American Society for Information Science, 50, 162-180
- Korfhage, R.R. (1997). Information storage and retrieval. New York: Wiley
- Landoni, M., & Bell, S. (2000). Information retrieval techniques for evaluating search engines: a critical overview. Aslib Proceedings, 52 (3), 124-129.
- Lawrence, S., Coetzee, F., Glover, E., Pennock, D., Flake, G., Nielsen, F., Krovetz, B., Kruger, A., & Giles, L. (2001). Persistence of Web references in scientific research. IEEE Computer, 34(2), 26-31.
- Lawrence, S., & Giles, C.L. (1998b). Searching the World Wide Web. Science, 280, 3, 98-100
- Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the web. Nature, 400, 107-109]
- Lawrence, S. & Giles, C.L. (2002). New Study on the Accessibility and Distribution

of Information on the Web. Available Online June, 19, 2002. http://www.neci.nec.com/~lawrence/searchtips.html

- Lynch, C. (1997). Searching the Internet. Scientific American. [On-line]. Available at http://www.sciam.com/0397issue/0397lynch.html
- Leighton, H.V., & Srivastava, J. (1999). First 20 precision among World Wide Web search services (search engines). Journal of the American Society for Information Science, 50, 870-881.
- Lempel, R., & Moran, S. (2000). The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In 9th International World Wide Web Conference (WWW9). [On-line]. Available at http://www9.org/w9cdrom/start.html.
- Meghabghab, G (2001) Discovering authorities and hubs in different topological Web graph structures. Information Processing and Management: an International Journal, Volume 38 , Issue 1, pp. 111-140
- Moukdad, H. (2002). Language Based Retrieval of Web Documents: an Analysis of Arabic Recognition Cabablities of Two Major Search Engines. In proceedings of the 65th ASISIT Annual Meeting, vol. 39 (ASISIT 2002), pp. 551-563
- Mowshowitz, A., Kawaguchi, A. (2002). Assessing Bias in Search Engines. Information Processing and Management, 35)4), pp. 443-462
- Najork, M., & Wiener, J.L. (2001). Breadth-first search crawling yields high-quality pages. In 10th International World Wide Web Conference (WWW10). [On-line]. Available at http://www10.org/cdrom/papers/208/
- Notess, G, R. (2004) Search Engines Statisitcis: Relative Size Showdown. Retrieved Agust, 2005 Fro http://www.searchengineshowdown.com/stats/size.shtml
- O’Meara, T., & Patel, A. (2001). A topic-specific Web robot model based on restless bandits. IEEE Internet Computing 5(2), 27-3
- Oppenheim, C., Morris, A., & McKnight, C. (2000). The evaluation of WWW search engines. Journal of Documentation 56, 190-211
- Qin, J., & Wesley, K. (1998). Web indexing with meta fields: a survey of Web objects in polymer chemistry. Information Technology and Libraries, 17, 149-156.
- Rasmussen, Edie (2003). Indexing and Retrieval for the Web. Annual Review of Information Science and Technology, vol. 37, Chapter 3, pp91-123
- Rousseau, R. (1998/9). Daily time series of common single word searches in AltaVista

and NorthernLight. Cybermetrics, 2/3, Issue 1, Paper 2. [On-line]. Available at http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.pdf

- Schwartz, C. (1998). Web search engines. Journal of the American Society for Information Science, 49, 973-982
- Selberg, E., & Etzioni, O. (2000). On the instability of Web search engines. Proceedings of the RIAO'2000 Conference. Paris: C.I.D. [On-line]. Available at http://133.23.229.11/~ysuzuki/Proceedingsall/RIAO2000/Wednesday/19plenary2.pdf
- Smith, Z. The Truth about Web: Crawling Towards eternity. Web Techniqu Magazine, May, 2005. Retrieved from the Web at 27, June, 2005 http://www.webtechnique.com/features/2005/05.
- Stanley, T. (1997b). Moving up the ranks. Ariadne [On-line], 12. Available http://www.ariadne.ac.uk/issue12/search-engines
- Su, L. (1997). Developing a comprehensive and systemic model of user evaluation of Web-based search engines. Proceedings of the 18th National Online Meeting (pp. 335-344). Medford, NJ: Information Today.
- Sullivan. D (2005a, December 11). Search engine sizes [On-line]. Available: http://www.searchenginewatch.com/reports/sizes.htm
- Sullivan, D (2002). How Search Engines Work. Retrieved from the Web at, June, 25, 2005
- Tomaiuolo, N.G., & Packer, J.G. (1996). An analysis of Internet search engines: assessment of over 200 search queries. Computers in Libraries, 16(6), 58-62.
- Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology. In S. Hardin (Ed.), Proceedings of the 59th Annual Meeting of the American Society for Information Science (pp. 127-135). Medford, NJ: American Society for Information Science
- Turner, T.P., & Brackbill, L. (1998). Rising to the top: evaluating the use of the HTML meta tag to improve retrieval of World Wide Web documents through Internet search engines. Library Resources and Technical Services, 42, 258-271.
- Vaughan, Liwen & Thelwall, Mike. (2004). Search Engines Bias: Evidence and Possible Causes, Information Processing and Mangement, vol. 40 no. 4, pp693-707.
- Voorhees, E.M. (2000a). Overview of the TREC-9 Question Answering Track. In The Ninth Text REtrieval Conference: TREC-9. (NIST Special Publication 500-249). [On-line]. Available at http://trec.nist.gov/pubs/trec9/papers/qa_overview.pdf
- Voorhees, E.M. (2000b). Report on TREC-9. SIGIR Forum, 34(2), 1-8.

- Wall, Aaron. Search Marketing. History of Search Engines & Web History. Retrieved from the WWW at May, 16, 2005. http://www.search-marketing.info/search-engine-history/
- Woodruff, A., Aoki, P.M., Brewer, E., Gauthier, P., & Rowe, L.A. (1996). An investigation of documents from the World Wide Web. In: Fifth International World Wide Web Conference (WWW5) [On-line]. Available at
- http://www5conf.inria.fr/fich_html/papers/P7/Overview.html
- (Also published as a special issue of Computer Networks and ISDN Systems, Volume 28 (7-11), 963-980
- Yuwono, B., & Lee, D.L. (1996). Search and ranking algorithms for locating resources on the World Wide Web. In: Proceedings of the 12th International Conference on Data Engineering (pp. 164-171). [On-line]. Available at http://www.cs.ust.hk/~dlee/.

* **خالد عبد الفتاح محمد**

مدير حلول المعرفة ومكتبة دبي الرقمية بمؤسسة محمد بن راشد ال مكتوم للمعرفة.

الجنسية: مصري.

الدرجة: أستاذ دكتور علوم المكتبات والمعلومات بجامعة الفيوم.

الجوائز:

- جائزة أكاديمية البحث العلمي في مجال المعلوماتية وإدارة المعرفة عام 2008.
- جائزة أفضل بحث في مؤتمر جمعية المكتبات المتخصصة فرع الخليج لعام 2009.
- جائزة التميز في النشر العلمي من جامعة الفيوم 2015.
- جائزة التميز في النشر العلمي من جامعة الفيوم عام 2016.

مشروعات التطوير:

- مكتبة دبي الرقمية وحلول معرفية بمؤسسة محمد بن راشد ال مكتوم لمعرفة.
- الاتحاد العربي للمكتبات الرقمية.
- بنك المعرفة المصري.
- أتحاد المكتبات الجامعية المصرية.
- الفهرس الموحد للمكتبات الجامعية المصرية.
- المستودع الرقمي للمكتبات الجامعية المصرية.
- التحول الرقمي بمعهد التخطيط القومي المصري.
- المكتبة الرقمية لجامعة بتسبرج بالولايات المتحدة.

صدرله

- الميتاداتا: أسسها النظرية وتطبيقاتها العملية، القاهرة: الدار المصرية اللبنانية للطباعة والنشر والتوزيع، 2013.
- الإنترنت: المكونات والتكنولوجيات والتطبيقات في المكتبات ومؤسسات المعلومات، مكتبة المتنبي، 2014.

Merging Multiple Search Results Approach for Meta Search Engines. VDM Publishing group, German, 2009.

- كيف تكتب بحثاً علمياً وتنشره، القاهرة، الدار المصرية اللبنانية ومؤسسة محمد بن راشد آل مكتوم، 2008.
- الويب الدلالي وتطبيقاته في المكتبات، السعودية، مكتبة المتنبي 2018.
- كما نشر أكثر من 25 بحثاً في دوريات علمية أجنبية وعربية ومؤتمرات علمية. يمكن مراجعة قائمة الأبحاث من خلال موقعه على جوجل العلمي:

https://scholar.google.ae/citations?hl=ar&user=NHpxNp8AAAAJ